SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

001

BICYCLES AND SPANNING TREES*

KENNETH A. BERMAN†

Abstract. Let G be a connected multigraph and let $(A, +, 0)$ be any Abelian group. For k an integer, let $A(k)$ denote the subgroup of A given by $A(k)=\{a\in A \mid ka=0\}$. A *bicycle* over A is a cycle over A that is also a cocycle. The set $B(A)$ of bicycles over A determines a group. In this paper we show that the spanning tree number t of G has a unique factorization $t=t_1t_2\cdots t_m$ such that t_i is a multiple of t_{i+1}, $i=1,2,\cdots,m-1$ and such that for every Abelian group A the group $B(A)$ of bicycles over A is isomorphic to $A(t_1)\times A(t_2)\times\cdots\times A(t_m)$. Using this result we obtain a number of results on the spanning tree number including two formulae for the spanning tree number.

1. Introduction and definitions. Let G be a connected multigraph with vertex set V and edge set E. All multigraphs considered in this paper will be without loops. A *tree* of G is a connected subgraph that has no circuits. A tree may consist of a single vertex. A *spanning tree* is a tree that spans the vertices. Throughout this paper t will denote the number of spanning trees.

Let $(A, +, 0)$ be an Abelian group. A *weighting of the vertices over* A is a mapping f from V into A. For $v\in V$, we will refer to $f(v)$ as the *weight of* f *on* v. Let $W_V(A)$ denote the set of vertex weightings over A. The set $W_V(A)$ determines a group where addition is given by

$$(f_1+f_2)(v)=f_1(v)+f_2(v) \tag{1.1}$$

for $f_1, f_2\in W_V(A)$ and $v\in V$. Analogously, we have a group $W_E(A)$ of edge weightings over A.

Now consider a commutative ring $(R, +, \cdot, 0, 1)$. (Note that $(R, +, 0)$ is an Abelian group.) The set $W_V(R)$ of vertex weightings over R determines a module over R where group addition is given by (1.1) and scalar multiplication is given by

$$(\lambda f)(v)=\lambda(f(v)) \tag{1.2}$$

$\lambda\in R, f\in W_V(R)$ and $v\in V$. Analogously, we have a module $W_E(R)$ of edge weightings over R. If F is a field then the set $W_V(F)$ of vertex weightings over F and the set $W_E(F)$ of edge weightings over F determine vector spaces over F.

Orient the edges of G arbitrarily. For $v\in V$, let $N^+(v)$ and $N^-(v)$ denote the set of edges having head v and tail v, respectively. A *cycle* over A is a weighting c of the edges such that for each vertex $v\in V$

$$\sum_{e^+\in N^+(v)} c(e^+)=\sum_{e^-\in N^-(v)} c(e^-). \tag{1.3}$$

The set $C(A)$ of cycles over A is a subgroup of $W_E(A)$.

For $e\in E$ let $h(e)$ and $t(e)$ denote the head and tail of edge e, respectively. The *coboundary operator* δ is the mapping from $W_V(A)$ to $W_E(A)$ given by

$$\delta f(e)=f(h(e))-f(t(e)) \tag{1.4}$$

for $f\in W_V(A)$ and $e\in E$. A *coboundary* or *cocycle* over A is a weighting y of the edges such that $y=\delta f$ for some $f\in W_V(A)$. The set $Y(A)$ of cocycles over A is a subgroup of $W_E(A)$.

Note that if we reverse the direction of an edge e and replace the weight on e with negative that weight then a cycle remains a cycle and a cocycle remains a cocycle.

* Received by the editors November 29, 1982, and in revised form October 3, 1984.

† Department of Mathematics, Wesleyan University, Middletown, Connecticut 06457.

Thus, in this sense, the definition of a cycle and a cocycle is independent of the orientation chosen. Cycles and cocycles are studied in electrical network theory [4]. Tutte [11] studied cycles and cocycles over the integers mod k in connection with face k-colorings of a plane graph and vertex k-colorings of a general graph.

A *bicycle* over A is an edge weighting over A that is both a cycle and a cocycle. The group $B(A)$ of bicycles over A is the intersection group of the cycle and cocycle groups over A, i.e., $B(A) = C(A) \cap Y(A)$. Bicycles, particularly over the field of integers mod 2, have been studied by a number of authors. See [3], [6], [7], [8], [9].

For k an integer, let $A(k)$ denote the subgroup of A given by $A(k) = \{a \in A \mid ka = 0\}$. In § 2, we show that the spanning tree number t of G has a unique factorization $t = t_1 t_2 \cdots t_m$ such that t_i is a multiple of t_{i+1}, $i = 1, 2, \cdots, m-1$ and such that for every Abelian group A the group $B(A)$ of bicycles over A is isomorphic to the direct product group $A(t_1) \times A(t_2) \times \cdots \times A(t_m)$. We call the factorization $t = t_1 t_2 \cdots t_m$ the *principal factorization* of t. This result has various implications. One immediate implication is the following existence theorem. There exists a nonzero bicycle over an Abelian group A if and only if A contains a nontrivial subgroup whose order divides the spanning tree number. The latter result generalizes an existence theorem of Shank (see [6]) on bicycles over a field.

In § 3, we obtain a formula for the factors in a principal factorization. In § 4, we show that a planar graph and its dual is an example of a pair of graphs that have the same spanning tree number with the same principal factorization.

Let $\mathbf{Z}_k$ denote the ring of integers modulo k. We will refer to a bicycle over $\mathbf{Z}_k$ as a *k-bicycle*. For convenience, we denote the module of k-bicycles by $B(k)$, i.e., $B(k) = B(\mathbf{Z}_k)$. In § 5, we show that the number $\beta(k)$ of k-bicycles is given by

$$\beta(k) = \prod_{i=1} \text{g.c.d.}\,(k, t_i) \tag{1.5}$$

where $t = t_1 t_2 \cdots t_m$ is the principal factorization of the spanning tree number t and where g.c.d. (x, y) denotes the greatest common divisor of the integers x and y. As an immediate corollary of this we have $\beta(k)$ divides t for all $k \geqq 2$ and $\beta(k) = t$ if k is a multiple of t. We use this to prove a number of divisibility results on the spanning tree number.

A k-bicycle b is *reducible* if the weight of b on each edge is divisible by d where d is a noninvertible element of the ring $\mathbf{Z}_k$. Otherwise b is *irreducible*. Let $\gamma(k)$ denote the number of irreducible k-bicycles and set $\gamma(1) = 1$. In § 6, we show that if there exists an irreducible k-bicycle then k is a divisor of t. Further, the total number of irreducible k-bicycles over all the positive integers k equals the spanning tree number, i.e.,

$$t = \sum_{k=1}^{\infty} \gamma(k). \tag{1.6}$$

Let p be a positive prime number and i a positive integer. The set $pB(p^i) = \{pb \mid b \in B(p^i)\}$ is a submodule of $B(p^i)$. Consider the quotient module

$$\bar{B}(p^i) = \frac{B(p^i)}{pB(p^i)}. \tag{1.7}$$

That is, $\bar{B}(p^i)$ consists of congruence classes of elements from $B(p^i)$ where two p^i-bicycles b_1 and b_2 are congruent if $b_1 - b_2 = pb_3$ for some p^i-bicycle b_3. The quotient module $\bar{B}(p^i)$ is a vector space over the field of integers mod p. We will refer to $\bar{B}(p^i)$ as the *quotient p^i-bicycle space*. Let $\rho(p^i)$ denote the dimension of $\bar{B}(p^i)$. In § 7, we show that the prime factorization of the spanning tree number t is given by

$$t = \prod_{p \in \mathcal{P}} p^{\rho(p)+\rho(p^2)+\rho(p^3)+\cdots} \tag{1.8}$$

where $\mathcal{P}$ denotes the set of positive prime numbers.

In § 8, we employ the above result to strengthen a theorem of Shank [8] on left-right paths and spanning trees in planar graphs.

2. Characterization theorem. Let G be a connected multigraph with vertex set $V = \{v_1, v_2, \cdots, v_n\}$ and edge set E. Let $(A, +, 0)$ be an Abelian group. In this section, we characterize the group $B(A)$ of bicycles over A.

Before stating the main theorem we discuss some preliminary results. For $v \in V$, let $N(v)$ denote the set of edges incident with vertex v and let $d(v)$ denote the degree of vertex v. For $e \in N(v)$, let v_e denote the end vertex of edge e different from v. A weighting f of the vertices over A is *balanced* if for every vertex v

$$d(v)f(v) = \sum_{e \in N(v)} f(v_e). \tag{2.1}$$

Let $L(A)$ denote the set of balanced vertex weightings over A. Then, $L(A)$ is a subgroup of $W_V(A)$. A vertex weighting is *constant* if the weight on every vertex is the same. Clearly a constant vertex weighting is balanced. The following proposition is a joint result (unpublished) of the author and H. Shank.

PROPOSITION 2.1. *The coboundary operator δ is a surjective homomorphism from the group $L(A)$ of balanced vertex weightings over A to the group $B(A)$ of bicycles over A whose kernel is the group of constant weightings over A.*

Proof. It is immediate that the kernel of δ is the group of constant vertex weightings. Let $f \in W_V(A)$. We wish to show that $g = \delta f$ is a cycle iff f is a balanced vertex weighting. Now g is a cycle iff for every vertex v

$$\begin{aligned}
&\sum_{e^+ \in N^+(v)} g(e^+) = \sum_{e^- \in N^-(v)} g(e^-) \\
&\Leftrightarrow \sum_{e^+ \in N^+(v)} (f(h(e^+)) - f(t(e^+))) = \sum_{e^- \in N^-(v)} (f(h(e^-)) - f(t(e^-))) \\
&\Leftrightarrow \sum_{e^+ \in N^+(v)} f(h(e^+)) + \sum_{e^- \in N^-(v)} f(t(e^-)) = \sum_{e^- \in N^-(v)} f(h(e^-)) + \sum_{e^+ \in N^+(v)} f(t(e^+)) \\
&\Leftrightarrow d(v)f(v) = \sum_{e \in N(v)} f(v_e).
\end{aligned}$$

The last equation is true iff f is a balanced vertex weighting.

COROLLARY 2.2. *The group $L(A)$ of balanced vertex weightings is isomorphic to the direct product of A and the group $B(A)$ of bicycles over A, i.e.,*

$$L(A) \cong A \times B(A). \tag{2.2}$$

The *Kirchhoff matrix* $\mathbf{K} = (k_{ij})$ is the $n \times n$ matrix such that $k_{ii} =$ (degree of vertex v_i) for $i \in \{1, 2, \cdots, n\}$ and $k_{ij} = -$(the number of edges joining vertices v_i and v_j) for $i, j \in \{1, 2, \cdots, n\}$, $i \neq j$ ($k_{ij} = 0$ if there are no edges joining v_i and v_j.). Let $\mathbf{K}[i{:}j]$ be the matrix obtained from the Kirchhoff matrix $\mathbf{K}$ by stroking out the ith row and the jth column and let $\det \mathbf{K}[i{:}j]$ denote the determinant of $\mathbf{K}[i{:}j]$. The *ij-cofactor* C_{ij} of the Kirchhoff matrix is given by $C_{ij} = -1^{i+j} \det \mathbf{K}[i{:}j]$. The following theorem is a classical result known as the matrix-tree theorem.

MATRIX-TREE THEOREM *The spanning tree number t of G is equal to any cofactor of the Kirchhoff matrix, i.e. $\forall i, j \in \{1, 2, \cdots, n\}$*

$$t = (-1)^{i+j} \det \mathbf{K}[i{:}j].$$

For references the matrix-tree theorem see [1], [4], [10].

Let $M=(m_{ij})$ be an $n\times n$ integer matrix. For A an Abelian group let A^n be the group obtained by taking the direct product of A with itself n times. Let $H(\mathbf{M}, A)$ denote the set of all $\underset{\sim}{x}=(x_1, x_2, \cdots, x_n)\in A^n$ satisfying the homogeneous equations

$$\sum_{j=1}^{m} m_{ij}x_j=0 \qquad (i=1,2,\cdots,n). \tag{2.3}$$

We will represent these equations in matrix notation by

$$\mathbf{M}\underset{\sim}{x}^t=0, \qquad \underset{\sim}{x}\in A^n \tag{2.4}$$

($\underset{\sim}{x}^t$ denotes the transpose of $\underset{\sim}{x}$). Clearly, $H(\mathbf{M}, A)$ is a subgroup of A^n. For $\underset{\sim}{x}=(x_1, x_2, \cdots, x_n)\in A^n$ let $\hat{x}$ denote the vertex weighting over A such that $\hat{x}(v_i)=x_i$, $i=1,2,\cdots,n$.

PROPOSITION 2.3. *Let* $\mathbf{K}$ *be the Kirchhoff matrix. Then the group* $L(A)$ *of balanced vertex weightings is isomorphic to the group* $H(\mathbf{K}, A)$, *i.e.*,

$$L(A)\cong H(\mathbf{K}, A). \tag{2.5}$$

Proof. It is immediate from the definitions of a balanced vertex weighting and the Kirchhoff matrix that $\mathbf{K}\underset{\sim}{x}^t=0$ if and only if $\hat{x}$ is a balanced vertex weighting.

We are now ready to state the main theorem. For k an integer, let $A(k)$ denote the subgroup of A given by $A(k)=\{a\in A\,|\,ka=0\}$.

THEOREM 2.4. *Let* G *be a connected multigraph with spanning tree number* t. *Then* t *has a unique factorization* $t=t_1t_2\cdots t_m$ *such that* t_i *is a multiple of* t_{i+1}, $i=1,2,\cdots,m-1$ *and such that for every Abelian group* A *the group* $B(A)$ *of bicycles over* A *is isomorphic to the direct product group* $A(t_1)\times A(t_2)\times\cdots\times A(t_m)$, *i.e.*,

$$B(A)\cong A(t_1)\times A(t_2)\times\cdots\times A(t_m). \tag{2.6}$$

Proof. We prove the theorem with the aid of two lemmas. Let $M=(m_{ij})$ be any $n\times n$ integer matrix. By a classical result (see [5]) there exist invertible $n\times n$ integer matrices $\mathbf{P}$ and $\mathbf{Q}$ (i.e., $\det\mathbf{P}=\pm1$ and $\det\mathbf{Q}=\pm1$) and a diagonal matrix $\mathbf{D}=\operatorname{diag}(d_1, d_2, \cdots, d_n)$ of nonnegative integers such that d_i divides d_{i+1}, $i=1,2,\cdots,n$ (by convention 0 divides 0) and such that

$$\mathbf{M}=\mathbf{PDQ}. \tag{2.7}$$

The integers $d_1, d_2, \cdots, d_n$ are the invariant factors of $\mathbf{M}$.

LEMMA 2.5. *Let* $\mathbf{M}$ *be an* $n\times n$ *integer matrix and let* $d_1, d_2, \cdots, d_r$ *be the nonzero invariant factors of* $\mathbf{M}$ *such that* d_i *divides* d_{i+1}, $i=1,2,\cdots,r-1$. *Then*

$$H(\mathbf{M}, A)\cong A^{n-r}\times A(d_1)\times A(d_2)\times\cdots\times A(d_r). \tag{2.8}$$

Proof. For $\underset{\sim}{x}\in A^n$ let $\underset{\sim}{x}'=\mathbf{Q}\underset{\sim}{x}^t$.

Now

$$\mathbf{M}x^t=0 \Leftrightarrow (\mathbf{PDQ})x^t=0 \Leftrightarrow \mathbf{D}\underset{\sim}{x}'=0.$$

The last step follows since $\mathbf{P}$ is invertible. Since $\mathbf{Q}$ is invertible it follows that $H(\mathbf{M}, A)$ and $H(\mathbf{D}, A)$ are isomorphic groups. It is immediate that $H(\mathbf{D}, A)=A(d_1)\times A(d_2)\times\cdots\times A(d_n)=A(d_1)\times A(d_2)\times\cdots\times A(d_r)\times A^{n-r}$. This proves the lemma.

LEMMA 2.6. *Let* $x_1, x_2, \cdots, x_r$ *be positive integers greater than* 1 *such that* x_i *is a multiple of* x_{i+1}, $i=1,2,\cdots,r-1$ *and let* $y_1, y_2, \cdots, y_s$ *be positive integers greater than* 1 *such that* y_i *is a multiple of* y_{i+1}, $i=1,2,\cdots,s-1$. *For* A *an Abelian group set*

$X(A)=A(x_1)\times A(x_2)\times\cdots\times A(x_r)$ and $Y(A)=A(y_1)\times A(y_2)\times\cdots\times A(y_s)$. Then $X(A)\cong Y(A)$ for every Abelian group A iff $r=s$ and $x_i=y_i$, $i=1,2,\cdots,r(=s)$.

Proof. If $r=s$ and $x_i=y_i$, $i=1,2,\cdots,r$ then trivially $X(A)=Y(A)$ for every Abelian group A. Conversely, suppose that $X(A)\cong Y(A)$ for every Abelian group A. Assume that $x_i\neq y_i$ for some index i. We show that a contradiction arises. Let p be a prime that divides x_r. Then p divides x_i for all $i<r$ since x_i is a multiple of x_r for all $i<r$. Consider the group $\mathbf{Z}_p$ of integers mod p. Clearly the order of $X(\mathbf{Z}_p)$ equals p^r. The prime p must divide y_s since otherwise we would have that the order of $Y(\mathbf{Z}_p)$ equals 1 contradicting that fact that $X(\mathbf{Z}_p)\cong Y(\mathbf{Z}_p)$. It follows that the order of $Y(\mathbf{Z}_p)$ equals p^s. Since the order of $X(\mathbf{Z}_p)$ must equal the order of $Y(\mathbf{Z}_p)$ we have that $r=s$. Now let j be the highest index such that $x_i\neq y_i$, i.e., $x_i=y_i$ for $i>j$ and $x_j\neq y_j$. Assume without loss of generality that $x_j>y_j$. Then there must exist a prime q such that the exponent ε of the highest power of q that divides x_j is strictly greater than the exponent ε' of the highest power of q that divides y_j. Consider the group $\mathbf{Z}_{q^\varepsilon}$ of the integers mod q^ε. It is easily shown that if A is the group $\mathbf{Z}_l$ of integers mod l then the order of $A(k)$ equals the greatest common divisor of k and l. Thus it follows that the order of $X(\mathbf{Z}_{q^\varepsilon})$ equals $\prod_{i=1}^r$ g.c.d. (x_i, q^ε) and the order of $Y(\mathbf{Z}_{q^\varepsilon})$ equals $\prod_{i=1}^r$ g.c.d. (y_i, q^ε). Now

$$\text{g.c.d.}\,(x_j, q^\varepsilon)=q^\varepsilon>q^{\varepsilon'}=\text{g.c.d.}\,(y_j, q^\varepsilon).$$

Since x_i is a multiple of x_j for $i<j$

$$\text{g.c.d.}\,(x_i, q^\varepsilon)=q^\varepsilon\geqq\text{g.c.d.}\,(y_i, q^\varepsilon)\quad\text{for } i<j.$$

Since by assumption $x_i=y_i$ for $i>j$ we have

$$\text{g.c.d.}\,(x_i, q^\varepsilon)=\text{g.c.d.}\,(y_i, q^\varepsilon)\quad\text{for } i>j.$$

Combining these inequalities we have

$$|X(\mathbf{Z}_q{}^\varepsilon)|=\prod_{i=1}^r \text{g.c.d.}\,(x_i, q^\varepsilon)>\prod_{i=1}^r \text{g.c.d.}\,(y_i, q^\varepsilon)=|Y(\mathbf{Z}_{q^\varepsilon})|$$

contradicting the fact that $X(\mathbf{Z}_{q^\varepsilon})\cong Y(\mathbf{Z}_{q^\varepsilon})$. This proves the lemma.

We now prove Theorem 2.4. Let $d_1, d_2,\cdots, d_n$ be the invariant factors of the Kirchhoff matrix $\mathbf{K}$. Since $\det\mathbf{K}=0$ it follows that $d_n=0$. By the matrix-tree theorem all the (first-order) cofactors equal the spanning tree number t. This implies that $t=d_1d_2\cdots d_{n-1}$. Employing Proposition 2.3 and Lemma 2.5 we have that

$$L(A)\cong H(\mathbf{K}, A)\cong A\times A(d_1)\times\cdots\times A(d_{n-1}),$$

and by Corollary 2.2 we have that

$$B(A)\cong A(d_1)\times A(d_2)\times\cdots\times A(d_{n-1}).$$

Now set $t_i=d_{n-i}$, $i=1,2,\cdots,n-1$ and let m be the largest index such that $t_i\neq 1$. Then $t=t_1t_2\cdots t_m$ and t_i is a multiple of t_{i+1}, $i=1,2,\cdots,m-1$. Also

$$B(A)\cong A(t_1)\times A(t_2)\times\cdots\times A(t_m).$$

The fact that the factorization $t=t_1t_2\cdots t_m$ is determined uniquely follows from Lemma 2.6.

This completes the proof of the theorem.

We will call the factorization $t=t_1t_2\cdots t_m$ the *principal factorization* of t.

COROLLARY 2.7. *Let G be a connected multigraph with spanning tree number t. Then there exists a bicycle over an Abelian group A iff A contains a nontrivial, finite subgroup whose order divides t.*

Proof. First suppose there exists a nonzero bicycle over A, i.e., $B(A)$ is nontrivial. Since $B(A) \cong A(t_1) \times A(t_2) \times \cdots \times A(t_m)$ where t_i is a multiple of t_{i+1}, $i = 1, 2, \cdots, m-1$ it follows that $A(t_1)$ is nontrivial. Let $a \in A(t_1)$, $a \neq 0$. By definition $t_1 a = 0$. Hence the order of the cyclic subgroup C of A generated by the element a divides t_1. Since t_1 divides t, the order of C divides t.

Conversely suppose A contains a finite subgroup S whose order divides t. Let $a \in S$, $a \neq 0$. Then $ta = 0$ or equivalent $(t_1 t_2 \cdots t_m)a = 0$. This implies that either $t_m a = 0$ or there exists a j, $1 \leqq j < m$ such that $t_j(t_{j+1} \cdots t_m a) = 0$ and $a' = t_{j+1} t_{j+2} \cdots t_m a \neq 0$. If the former case is true then $A(t_m)$ is nontrivial since it contains the nonzero element a, and if the latter case is true then $A(t_j)$ is nontrivial since it contains the nonzero element a'. In either case Theorem 2.4 implies that $B(A)$ is nontrivial. Hence there exists a nonzero bicycle over A.

Corollary 2.6 generalizes the following result of Shank (see [6]).

COROLLARY 2.8 (Shank). *Let G be a connected multigraph having spanning tree number t. Then there exists a bicycle over a field F iff the characteristic of F is nonzero and divides t.*

Proof. Let p denote the characteristic of F. If $p = 0$ then F cannot contain a finite additive subgroup other than the trivial subgroup and hence by Corollary 2.6 there is no bicycle over F. Suppose $p \neq 0$ and p divides t. Clearly p is the order of the subfield (subgroup) generated by the multiplicative identity. By Corollary 2.6 this implies that there exists a bicycle over F. Conversely suppose there exists a bicycle over F. Then by Corollary 2.6 F contains a nontrivial additive subgroup whose order divides t. But the characteristic p divides the order of any additive subgroup of F. Hence p divides t.

The special case of Corollary 2.7 when $p = 2$ was discovered independently by Chen [3].

3. Formula for the factors in a principal factorization. Assign a linear ordering $<$ to the vertex set V of the multigraph G. For k a positive integer let $\mathcal{V}_k$ denote the collection of all sets of k vertices. Consider any two sets $R = \{r_1, r_2, \cdots, r_k\}$ and $S = \{s_1, s_2, \cdots, s_k\}$ from $\mathcal{V}_k$ where $r_i < r_j$ and $s_i < s_j$ for $i < j$. Let $\mathcal{P}_k$ denote the set of all permutations of $\{1, 2, \cdots, k\}$ and suppose $\sigma \in \mathcal{P}_k$. An *(R, S, σ)-forest* is a set of k vertex disjoint trees $T_1, T_2, \cdots, T_k$ whose union spans the vertices of G such that tree T_i contains the vertices r_i and $s_{\sigma(i)}$; $i = 1, 2, \cdots, k$. Let $f_\sigma(R, S)$ denote the number of (R, S, σ)-forests and set

$$f_k(R, S) = \sum_{\sigma \in \mathcal{P}_k} \operatorname{sign} \sigma f_\sigma(R, S). \tag{3.1}$$

Now let f_k denote the greatest common divisor of $f_k(R, S)$ over all sets $R, S \in \mathcal{V}_k$, i.e.,

$$f_k = \text{g.c.d.}\,\{f_k(R, S) \mid R, S \in \mathcal{V}_k\}. \tag{3.2}$$

THEOREM 3.1. *Let G be a connected multigraph on n vertices whose spanning tree number t has principal factorization $t = t_1 t_2 \cdots t_m$. Set $t_k = 1$ for $k > m$ and let f_k be defined by* (3.2). *Then*

$$t_i = \frac{f_i}{f_{i+1}}, \qquad i = 1, 2, \cdots, n-1. \tag{3.3}$$

Proof. Let $d_1, d_2, \cdots, d_{n-1}$ be the nonzero invariant factors of the Kirchhoff matrix $\mathbf{K}$. Then $t_i = d_{n-i}$, $i = 1, 2, \cdots, n-1$. Let Δ_i denote the greatest common divisor over all the minors of $\mathbf{K}$ of size i. Set $\Delta_0 = 1$. Then by a result in [5] we have $d_i = \Delta_i / \Delta_{i-1}$,

$i=1,2,\cdots,n-1$. Thus

$$t_i=\frac{\Delta_{n-i}}{\Delta_{n-i-1}},\qquad i=1,2,\cdots,n-1. \tag{3.4}$$

For $R, S\in \mathcal{V}_i$ let $\mathbf{K}[R:S]$ be the submatrix obtained from the Kirchhoff matrix by stroking out the rows corresponding to the vertices in R and the columns corresponding to the vertices in S. Then by the all minors matrix-tree theorem (This is a stronger version of the matrix-tree theorem which gives a formula in terms of spanning forests for the minors of the Kirchhoff matrix. See [2].) we have

$$f_i(R,S)=\pm\det \mathbf{K}[R:S]. \tag{3.5}$$

(The actual sign preceding $\det \mathbf{K}[R:S]$ is given in [2] but we omit it here since it is not needed in our proof.) This implies that

$$f_i=\Delta_{n-i},\qquad i=1,2,\cdots,n-1. \tag{3.6}$$

Equation (3.3) of Theorem 3.1 follows from (3.4) and (3.6).

An *R-forest* is an (R, R, σ)-forest where σ is the identity permutation.

COROLLARY 3.2. *Let F be a connected multigraph whose spanning tree number t has principal factorization $t=t_1t_2\cdots t_m$. Then for any set R of k vertices, $k\leqq m$, the number f_R of R-forests is divisible by $t_kt_{k+1}\cdots t_m$.*

Proof. Observe that $f_\sigma(R,R)=0$ unless σ is the identity permutation in which case $f_\sigma(R,R)=f_R$. This implies that $f_k(R,R)=f_R$. It follows from the definition of f_k that f_k divides $f_k(R,R)$. Now $f_k=(f_k/f_{k+1})(f_{k+1}/f_{k+2})\cdots(f_{n-1}/f_n)$ since $f_n=1$ ($f_n=f_V=1$ since there is only one forest having n trees, namely the forest such that each tree is a vertex of G). Therefore by Theorem 3.1 we have that $f_k=t_kt_{k+1}\cdots t_m$. Hence $t_kt_{k+1}\cdots t_m$ divides f_R as stated in the corollary.

4. Graphs with the same principal factorization of t. It is easy to find examples of graphs that have the same spanning tree number but different principal factorizations. For example the complete graph on 4 vertices, the graph consisting of a circuit of length 16 and the graph on 7 vertices consisting of two circuits of length 4 having exactly one vertex in common each have 16 spanning trees with principal factorizations $16=8\cdot 2$, $16=16$ and $16=4\cdot 4$ respectively.

The following proposition gives natural examples of pairs of graphs having the same spanning tree number with the same principal factorization.

PROPOSITION 4.1. *A planar graph G and its dual G^d have the same spanning tree number with the same principal factorization.*

Proof. Embed G in the plane. To obtain dual graph G^d from G we place a vertex in every face of G and join two vertices u and v of G^d with an edge e' whenever the face f_u of G containing vertex u and the face f_v of G containing vertex v share a common edge e. Now orient the edges of G at random. Assume without loss of generality that the face f_u is on the left and the face f_v is on the right when travelling along edge e in the assigned direction. Orient edge e' of G^d so that it is directed from u to v. Now let A be any Abelian group and let f be an edge weighting of G over A. Consider the edge weighting f' of G^d given by $f'(e')=f(e)$, $e\in E(G)$. It is a simple exercise to show that f is a bicycle of G if and only if f' is a bicycle of G^d. It follows that the group $B_G(A)$ of bicycles over A in G and the group $B_{G^d}(A)$ of bicycles over A in G^d are isomorphic for every Abelian group A. Therefore, by Theorem 2.4, G and G^d have the same spanning tree number with the same principal factorization.

5. Module of k-bicycles and divisibility results on the spanning tree number. Let G be a connected multigraph and let $B(K)$ denote the module of k-bicycles (i.e., the module of bicycles over the ring $\mathbf{Z}_k$ of integers mod k).

THEOREM 5.1. *Let G be a connected multigraph with spanning tree number t having principal factorization $t=t_1t_2\cdots t_m$. Then the number $\beta(k)$ of k-bicycles is given by*

$$\beta(k)=\prod_{i=1}^{m}\text{g.c.d.}\,(k,t_i). \tag{5.1}$$

Proof. Let $A=\mathbf{Z}_k$. It is easily verified that for l an integer $|A(l)|=\text{g.c.d.}\,(k,l)$. By Theorem 2.4 we have that

$$B(A)\cong A(t_1)\times A(t_2)\times\cdots\times A(t_m).$$

Hence

$$\begin{aligned}\beta(k)=|B(A)|&=|A(t_1)|\,|A(t_2)|\cdots|A(t_m)|\\&=\prod_{i=1}^{m}\text{g.c.d.}\,(k,t_i).\end{aligned}$$

This proves the theorem.

The following corollary is immediate.

COROLLARY 5.2. *The number $\beta(k)$ of k-bicycles divides the spanning tree number t for all $k\geqq 2$. Further if k is a multiple of t then $\beta(k)=t$.*

COROLLARY 5.3. *For any $k\geqq 2$, the cardinality of any submodule M of the module $B(k)$ of k-bicycles divides the spanning tree number t.*

Proof. By Lagrange's theorem the order (cardinality) of any subgroup of a finite group divides the order of that group. Since a module is an additive group it follows that the cardinality of any submodule of a finite module divides the cardinality of that module. Hence the cardinality of a submodule M of $B(k)$ divides $\beta(k)=|B(k)|$. But $\beta(k)$ divides t by Corollary 5.2.

Corollary 5.3 is useful in obtaining divisibility results on the spanning tree number as will be demonstrated in the proofs of the following three theorems.

A graph G is *strict* if it has no multiple edges. The complement graph G^c of G is the graph whose edges join precisely those pairs of vertices not joined in G.

THEOREM 5.4. *Let G be a strict graph on n vertices whose complement graph G^c is disconnected have κ connected components. Then the spanning tree number of G is divisible by $n^{\kappa-2}$.*

Proof. Let $H_1,H_2,\cdots,H_\kappa$ be the κ connected components of G^c and let U_i be the vertices of G that belong to H_i, $i\in\{1,2,\cdots,\kappa\}$. Let n_i denote the cardinality of U_i, $i\in\{1,2,\cdots,\kappa\}$. Then $n=n_1+n_2+\cdots+n_\kappa$. Consider a vertex n-weighting f of G such that $f(v)=x_i$ for all $v\in U_i$, $i\in\{1,2,\cdots,\kappa\}$ where $x_i\in\mathbf{Z}_n$ and

$$n_1x_1+n_2x_2+\cdots+n_kx_k=0. \tag{5.2}$$

Then f is a balanced n-weighting since equation (5.2) implies that, for each $i\in\{1,2,\cdots,\}$,

$$(n-n_i)x_i=n_1x_1+\cdots+n_{i-1}x_{i-1}+n_{i+1}x_{i+1}+\cdots+n_kx_k \tag{5.3}$$

and a vertex $v\in U_i$ has weight x_i and is joined to n_j vertices having weight x_j (i.e., the vertices in U_j) $i,j=\{1,2,\cdots,\kappa\}$, $i\neq j$. Let M denote the submodule of $L(n)$ (the module of balanced vertex n-weightings) consisting of all balanced vertex n-weightings f obtained in this fashion. Let M' denote the submodule of $B(n)$ consisting of the

n-bicycles $\{\delta f \mid f \in M\}$. Clearly $|M| = n|M'|$. Now $|M|$ is divisible by n^{k-1} because any solution of (5.2) can be obtained by choosing $x_2, x_3, \cdots, x_n$ arbitrarily and solving for x_1. Hence $n^{\kappa-2}$ divides $|M'|$. By Corollary 5.3 we have that $n^{\kappa-2}$ divides t.

THEOREM 5.5. *Let G be a multigraph on n vertices whose edges can be partitioned into q cliques, $q < n$, such that the size of each clique is a multiple of r for some integer $r \geqq 2$. Then the spanning tree number t of G is divisible by r^{n-q-1}.*

Proof. Let $Q_1, Q_2, \cdots, Q_q$ be the q cliques of G and let $V = \{v_1, v_2, \cdots, v_n\}$ be the vertex set of G. Let $\mathbf{C} = (c_{ij})_{q\times n}$ be the vertex-clique incidence matrix, i.e.,

$$c_{ij} = \begin{cases} 1, & v_j \text{ belongs to clique } Q_i, \\ 0, & v_j \text{ does not belong to clique } Q_i. \end{cases} \tag{5.4}$$

Consider the module $H(\mathbf{C}, \mathbf{Z}_r)$, i.e., $H(\mathbf{C}, \mathbf{Z}_r) = \{\underset{\sim}{x} \in \mathbf{Z}_r^n \mid \mathbf{C}\underset{\sim}{x}^t = 0\}$. Since $\mathbf{C}$ is a $q \times n$ matrix it follows that r^{n-q} divides $|H(\mathbf{C}, \mathbf{Z}_r)|$. Let $\underset{\sim}{x} = (x_1, x_2, \cdots, x_n)$ be an element from $H(\mathbf{C}, \mathbf{Z}_r)$ and consider the vertex r-weighting f such that $f(v_j) = x_j$ for all vertices $v_j \in V$. It is easily verified that f is a balanced vertex r-weighting. The set M of all balanced vertex r-weightings f obtained in this way is a submodule of $L(r)$, the module of all balanced vertex r-weightings. Let M' denote the submodule of the module $B(r)$ of r-bicycles given by $M' = \{\delta f \mid f \in M\}$. Then $|M'| = 1/r|M| = 1/r|H(\mathbf{C}, \mathbf{Z}_r)|$. Therefore r^{n-q-1} divides $|M'|$. By Corollary 5.3 we have that r^{n-q-1} divides t.

The *line graph* $L(G)$ of a graph G is obtained by associating a vertex of $L(G)$ with each edge of G and joining two vertices of $L(G)$ whenever the corresponding edges of G are incident. An *r-regular* graph is a graph in which every vertex has degree r.

COROLLARY 5.6. *Let G be an r-regular graph on n edges having line graph $L(G)$. Then the spanning tree number of $L(G)$ is divisible by $r^{((r-2)/r)n-1}$.*

Proof. Since G is r-regular and has n-edges it has $2n/r$ vertices. This implies that $L(G)$ has n vertices whose edges can be partitioned into $2n/r$ cliques each of size r. Corollary 5.6 can now be immediately induced from Theorem 5.5.

Let G be a connected multigraph with vertices $v_1, v_2, \cdots, v_n$ and let l_{ij} denote the number of edges linking vertices v_i and v_j; $i, j \in \{1, 2, \cdots, n\}$. We will say that a multigraph H is *divisible* by G if the vertices of H can be partitioned into n classes $U_1, U_2, \cdots, U_n$ such that for $i, j \in \{1, 2, \cdots, n\}$ a vertex v in U_i either (1) is joined only to vertices in U_i or (2) for every $j \neq i$ is joined to exactly λ_{ij} vertices of U_j (and any number of vertices in U_i).

THEOREM 5.7. *If a connected multigraph H is divisible by a connected multigraph G then the spanning tree number t_H of H is divisible by the spanning tree number t_G of G.*

Proof. Let $L_{ij}(H)$ denote the set of edges in H having one end vertex in U_i and the other in U_j; $i, j \in \{1, 2, \cdots, n\}$. For $e \in L_{ij}(H)$, $i \neq j$, let $\hat{e}$ be any edge of G linking vertices v_i and v_j. Now orient the edges of G such that all the edges joining the same two vertices are directed the same. Orient the edges of H such that e has tail in U_i and head in U_j if $\hat{e}$ has tail v_i and head v_j. If both ends of e belong in the same class orient e arbitrarily. For $k \geqq 2$ let b_G be any k-bicycle of G. It is easily verified that b_H is a k-bicycle of H where for e an edge of H

$$b_H(e) = \begin{cases} 0, & e \in L_{ii} \text{ for some } i \in \{1, 2, \cdots, n\}, \\ b_G(\hat{e}), & e \in L_{ij} \text{ for some } i, j \in \{1, 2, \cdots, n\}, \quad i \neq j. \end{cases} \tag{5.5}$$

Let $M_H(k)$ denote the submodule of the k-bicycle module $B_H(k)$ of H consisting of all k-bicycles b_H which correspond to a k-bicycle b_G of G in the above fashion. Then $|M_H(k)| = |B_G(k)|$ (where $B_G(k)$ denotes the k-bicycle module of G). In particular

$|M_H(t_G)| = |B_G(t_G)|$. But $|B_G(t_G)| = t_G$ by Corollary 5.2. By Corollary 5.3 we have that $|M_H(t_G)|$ divides t_H. Hence t_G divides t_H.

6. Irreducible bicycles. A nonzero k-bicycle b is *reducible* if there exists a noninvertible, nonzero element $d \in \mathbf{Z}_k$ such that d divides $b(e)$ for all edges e. A k-bicycle is *irreducible* if it is nonzero and not reducible.

THEOREM 6.1. *Let G be a connected multigraph with spanning tree number t. If there exists an irreducible k-bicycle then k is a divisor of t.*

Proof. Suppose G has an irreducible k-bicycle b. Then the set of all scalar multiples of b is a submodule of the k-bicycle module $B(k)$ that contains k elements. Hence the theorem follows from Corollary 5.3.

It is easy to find examples where the converse of Theorem 6.1 does not hold when k is a composite number. Note that Shank's result (Corollary 2.7) implies that the converse holds when k is a prime number.

Let $\gamma(k)$ denote the number of irreducible k-bicycles. Set $\gamma(1)=1$. Let $D(k)$ denote the set of divisors of k. The following proposition relates the number of bicycles to the number of irreducible bicycles.

PROPOSITION 6.2. *For k a positive integer, $k \geqq 2$ let $\beta(k)$ and $\gamma(k)$ denote the number of k-bicycles and number of irreducible k-bicycles respectively. Then*

$$\beta(k) = \sum_{d \in D(k)} \gamma(d) \tag{6.1}$$

where the summation is over all divisors d of k.

Proof. Let b be a nonzero k-bicycle. Let k' be the largest positive integer that divides $b(e)$ for all edges e and let $d = k/k'$. With the k-bicycle b we associate the irreducible d-bicycle b' defined by $b'(e) = b(e)/k'$ for e an edge of G. (In defining b' we are using the fact that for d a divisor of k the module $\mathbf{Z}_d$ may be regarded as a submodule of $\mathbf{Z}_k$.) This determines a bijection between the k-bicycles and the irreducible d-bicycles where d divides k, $d \neq 1$. The zero k-bicycle is counted by $\gamma(1)$.

PROPOSITION 6.3. *For r and s two relatively prime positive integers*

$$\gamma(rs) = \gamma(r)\gamma(s). \tag{6.2}$$

Proof. Since $Z_{rs} \cong Z_r \oplus Z_s$ (where $\oplus$ denotes the direct sum of two rings) we have that $B(rs) \cong B(r) \oplus B(s)$ (where $\oplus$ denotes the direct sum of two modules. The underlying set is $B(r) \times B(s)$, i.e., $b \in B(r) \oplus B(s) \Leftrightarrow b(e) = b_r(e) \times b_s(e)$, $e \in E$ for some $b_r \in B(r)$ and $b_s \in B(s)$.) Proposition 6.3 follows from the fact that the direct product $b_r \times b_s$ corresponds to an irreducible bicycle in $B(rs)$ if and only if b_r is an irreducible bicycle in $B(r)$ and b_s is an irreducible bicycle in $B(s)$.

We now state the main theorem of this section.

THEOREM 6.4. *Let G be a connected multigraph with spanning tree number t. Then, the total number over all the positive integers k, of irreducible k-bicycles equals t, i.e.,*

$$t = \sum_{k=1}^{\infty} \gamma(k). \tag{6.3}$$

Proof. By Theorem 6.1, $\gamma(k)=0$ for every k that is not a divisor of t. Hence by Proposition 6.2 we have that $\beta(t) = \sum_{d \in D(t)} \gamma(d) = \sum_{k=1}^{\infty} \gamma(k)$. But by Corollary 5.2 we have that $\beta(t) = t$. This proves Theorem 6.4.

7. Prime factorization formula for the spanning tree number. Let p be a positive prime number and i a positive integer. The set $pB(p^i) = \{pb \mid b \in B(p^i)\}$ is a submodule of $B(p^i)$. Consider the quotient module

$$\bar{B}(p^i)=\frac{B(p^i)}{pB(p^i)}.$$

That is, $\bar{B}(p^i)$ consists of congruence classes of elements from $B(p^i)$ where two p^i-bicycles b_1 and b_2 are congruent if $b_1-b_2=pb_3$ for some p^i-bicycle b_3. The quotient module $\bar{B}(p^i)$ is a vector space over the field of integers mod p. We will refer to $\bar{B}(p^i)$ as the *quotient p^i-bicycle space.* Let $\rho(p^i)$ denote the dimension of $\bar{B}(p^i)$. Let $\varepsilon_p(t)$ denote the exponent of the highest power of p that divides the spanning tree number t.

PROPOSITION 7.1. *Let G be a connected multigraph with spanning tree number t. Then $\rho(p^i)\geqq\rho(p^j)$ for $i<j$. Further $\rho(p^i)=0$ for all $i>\varepsilon_p(t)$.*

Proof. Let i and j be two positive integers such that $i<j$. Since $\bar{B}(p^i)$ is a vector space over $\mathbf{Z}_p$ the cardinality of $\bar{B}(p^i)$ equals $p^{\rho(p^i)}$, i.e., $|\bar{B}(p^i)|=p^{\rho(p^i)}$. Similarly $|\bar{B}(p^j)|=p^{\rho(p^j)}$. With every p^j-bicycle b we may associate a p^i-bicycle b' given by $b'(e)=b(e)(\text{mod } p^i)$ for e an edge of G. It is immediate that the mapping taking $\bar{b}$ to $\bar{b'}$ is an injective mapping from $\bar{B}(p^j)$ to $\bar{B}(p^i)$. This implies that $|\bar{B}(p^i)|\geqq|\bar{B}(p^j)|$. Hence $\rho(p^i)\geqq\rho(p^j)$. By Theorem 5.1 there are no irreducible p^i-bicycles for any $i>\varepsilon_p(t)$. This implies that $\bar{B}(p^i)$ contains only the zero bicycle. Thus $\rho(p^i)=0$ for all $i>\varepsilon_p(t)$. This proves Proposition 7.1.

We now state that main theorem of this section which gives a formula for the prime factorization of the spanning tree number.

THEOREM 7.2. *Let G be a connected multigraph with t spanning trees. For p a positive prime number and i a positive number let $\rho(p^i)$ denote the dimension of the quotient p^i-bicycle space. Then the prime factorization of t is given by*

$$t=\prod_{p\in P} p^{\rho(p)+\rho(p^2)+\cdots+\rho(p^i)+\cdots} \tag{7.1}$$

where the product is over the set P of all positive prime numbers. (The product over an infinite number of one's is one.)

Proof. By Proposition 6.2

$$\beta(p^i)=1+\gamma(p)+\gamma(p^2)+\cdots+\gamma(p^i). \tag{7.2}$$

Now $|\bar{B}(p^i)|=p^{\rho(p^i)}$. Also, it is clear that $|\bar{B}(p^i)|=\beta(p^i)/\beta(p^{i-1})$. Hence $\beta(p^i)=p^{\rho(p^i)}\beta(p^{i-1})$. By induction we have that

$$\beta(p^i)=p^{\rho(p)+\rho(p^2)+\cdots+\rho(p^i)}. \tag{7.3}$$

Combining (7.2) and (7.3) we obtain

$$p^{\rho(p)+\rho(p^2)+\cdots+\rho(p^i)+\cdots}=1+\gamma(p)+\gamma(p^2)+\cdots+\gamma(p^i)+\cdots. \tag{7.4}$$

Employing Proposition 6.3 and Theorem 6.4 we have

$$\prod_{p\in P} p^{\rho(p)+\rho(p^2)+\cdots+\rho(p^i)+\cdots}=\prod_{p\in P}(1+\gamma(p)+\gamma(p^2)+\cdots+\gamma(p^i)+\cdots)$$

$$=\sum_{k=1}^{\infty}\gamma(k)=t.$$

This proves Theorem 7.2.

COROLLARY 7.3. *Let G be a connected multigraph with t spanning trees. For p a positive prime number and i a positive integer let $\rho(p^i)$ denote the dimension of the quotient p^i-bicycle space. Then, p^i divides t iff $\rho(p)+\rho(p^2)+\cdots+\rho(p^i)\geqq i$. Further $(p^i)^{\rho(p^i)}$ divides t.*

Proof. If $\rho(p)+\rho(p^2)+\cdots+\rho(p^i)\geqq i$ then Theorem 7.2 implies that p^i divides t. Conversely suppose p^i divides t. Assume $\rho(p)+\rho(p^2)+\cdots+\rho(p^i)<i$. Since by Proposition 7.1 we have that $\rho(p^j)\geqq\rho(p^i)$ for all $j<i$ it follows that $\rho(p^i)=0$. Again applying Proposition 7.1 we have that $\rho(p^j)=0$ for all $j>i$. Since p^i divides t Theorem 7.2 yields

$$\begin{aligned} i &\leqq \rho(p)+\rho(p^2)+\cdots+\rho(p^i)+\cdots \\ &= \rho(p)+\rho(p^2)+\cdots+\rho(p^i). \end{aligned}$$

This is a contradiction. Hence $\rho(p)+\rho(p^2)+\cdots+\rho(p^i)\geqq i$. This proves the first part of the Corollary.

Since by Proposition 7.1, $\rho(p^i)\leqq\rho(p^j)$ for all $j<i$ we have that

$$i\rho(p^i)\leqq\rho(p)+\rho(p^2)+\cdots+\rho(p^i).$$

But by Theorem 7.2 we have that $p^{\rho(p)+\rho(p^2)+\cdots+\rho(p^i)}$ divides t. Hence $p^{i\rho(p^i)}$ divides t.

8. 2-bicycles in planar graphs. Let G be a planar, connected multigraph embedded in the plane. In [9], Shank discovered a very simple way of determining the dimension $\rho(2)$ of the 2-bicycle space. A *left-right path* of G is a closed path such that the edge chosen at each vertex is alternatively the leftmost (labelled L) and rightmost (labelled R) edge. An edge may be transversed twice as long as it is not transversed in the same direction with the same label.

THEOREM 8.1. (Shank). *Let G be a planar, connected multigraph with L left-right paths. Then the dimension $\rho(2)$ of the 2-bicycle space is given by*

$$\rho(2)=L-1.$$

Shank employed this theorem to prove that the number t of spanning trees of G is odd if and only if G has exactly one left-right path. The following theorem which is an immediate consequence of Theorem 8.1 and Corollary 7.3 strengthens this result.

THEOREM 8.2. *Let G be a planar, connected multigraph having t spanning trees and L left-right paths. If there is exactly one left-right path then t is odd. Otherwise 2^{L-1} divides t.*

Acknowledgments. I would like to thank J. D. Reid and H. Shank for helpful comments.

REFERENCES

[1] NORMAN BIGGS, *Algebraic Graph Theory*, Cambridge Univ. Press, Cambridge, 1974.

[2] SETH CHAIKEN, *A combinatorial proof of the All Minors Matrix-Tree Theorem*, this Journal, 3 (1982), pp. 319-329.

[3] W. K. CHEN, *On vector spaces associated with a graph*, SIAM J. Appl. Math., 20 (1971), pp. 526-529.

[4] ———, *Applied Graph Theory, Graphs and Electrical Networks*, 2nd ed., North-Holland, New York, 1976.

[5] NATHAN JACOBSON, *Basic Algebra* 1, W. H. Freeman, San Francisco, 1974.

[6] STEPHEN B. MAURER., *Matrix generalizations of some theorems, on trees, cycles and cocycles in graphs*, SIAM J. Appl. Math., 30 (1976), pp. 143-148.

[7] R. C. READ AND P. ROSENSTIEHL, *On the principle edge tripartition of a graph*, Ann. Discrete Math., 3 (1978), pp. 195-226.

[8] H. SHANK, *Graph property recognition machines*, Math. Systems Theory, 5 (1971), pp. 45-49.

[9] ———, *The theory of left-right paths* in Combinatorial Mathematics III, Lecture Notes in Mathematics 452, Springer, Berlin, 1975, pp. 42-54.

[10] H. M. TRENT, *A note on the enumeration and listing of all maximal trees of a connected linear graph*, Proc. Nat. Acad. Sci. USA, 40 (1954), pp. 1004-1007.

[11] W. T. TUTTE, *On the imbedding of linear graphs in surfaces*, Proc. London Math. Soc., Ser. 2, 51 (1949), pp. 474-483.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

002

TELEPHONE PROBLEMS WITH FAILURES*

KENNETH A. BERMAN† AND MICHAEL HAWRYLYCZ†

Abstract. Consider a multigraph G on n vertices whose edges are linearly ordered. The vertices of G may represent people and the edges two-way communication between pairs of people. A vertex v is *k-failure-safe* in communicating with a vertex w if there is a path of ascending edges from v to w even when any k edges of G are deleted. In this paper, we show that the minimum size $\mu(n, k)$ of G such that one vertex communicates k-failure-safe with every other vertex is given by $\mu(n, k) = \lceil((k+2)/2)(n-1)\rceil$ for $k \leqq n-2$ and $\mu(n, k) = \lceil((k+1)/2)n\rceil$ for $k \geqq n-2$. We also show that for $k \geqq 1$ the minimum size $\tau(n, k)$ of G such that every vertex communicates k-failure-safe with every other vertex satisfies $\mu(n, k) + n - 2\lceil\sqrt{n}\rceil \leqq \tau(n, k) \leqq \lfloor(k+3/2)(n-1)\rfloor$. The problem of finding $\tau(n, k)$ for $k = 0$ is the well-known telephone problem.

1. Introduction. Consider a multigraph G with vertex set V and edge set E where E has been assigned a linear order. We will call such a multigraph a *communication network.* An *ascending path* from a vertex v to a vertex w is a path from v to w such that for any two edges of the path the edge closer to v is smaller in the linear order. A vertex v *communicates* with a vertex w if there is an ascending path from v to w (note that if v communicates with w this does not necessarily mean that w communicates with v). A vertex v communicates *k-failure-safe* with a vertex w if there is an ascending path from v to w even when any k edges of G are deleted.

One model of a communication network is a group of people who have made a sequence of telephone calls. The people are represented by vertices and the calls between pairs of people are represented by edges. The edge corresponding to the ith call occurs ith in the linear ordering of the edges. When a call is made, the two people exchange all their information. An ascending path from a person P_i to a person P_j indicates that P_j has received P_i's information. If P_i communicates k-failure-safe with P_j then P_j is guaranteed to know P_i's information even if there is the possibility that in up to k of the calls information is not exchanged.

Consider a communication network G on n vertices where one vertex v communicates k-failure-safe with every other vertex. Note that by reversing the linear order of the edges of G we obtain a communication network in which every vertex communicates k-failure-safe with v. Let $\mu(n, k)$ be the minimum number of edges in such a network. In this paper we show that $\mu(n, k) = \lceil((k+2)/2)(n-1)\rceil$ for $k \leqq n-2$ and $\mu(n, k) = \lceil((k+1)/2)n\rceil$ for $k \geqq n-2$.

A k-failure-safe *total* communication network is a communication network in which every vertex communicates k-failure-safe with every other vertex. Let $\tau(n, k)$ be the minimum number of edges in a k-failure-safe total communication network on n vertices. In § 3 we show that for $k \geqq 1$, $\tau(n, k)$ satisfies $\mu(n, k) + n - 2\lceil\sqrt{n}\rceil \leqq \tau(n, k) \leqq \lfloor(k+3/2)(n-1)\rfloor$. The telephone problem which was proposed by A. Boyd and solved by a number of authors is equivalent to finding $\tau(n, k)$ for $k = 0$. For references see [1], [2], [3], [4], [5], [6]. It is well-known that $\tau(n, 0) = 2n - 4$.

In § 4 we consider communication networks in which the edges are directed, the orientation of an edge indicating the direction in which information is passed. One model of this is a group of people who send telegraph messages to other people in the group. In a directed communication network a vertex v communicates k-failure-safe with a vertex w if there is an ascending *directed* path from v to w even when any k

* Received by the editors November 15, 1983, and in revised form August 1, 1984.

† Department of Mathematics, Wesleyan University, Middletown, Connecticut 06457.

arcs are deleted. It is shown that the minimum number $\vec{\mu}(n, k)$ of arcs in a directed communication network in which one vertex communicates k-failure-safe with every other vertex is given by $\vec{\mu}(n, k) = (k+1)(n-1)$. Further we show that the minimum number $\vec{\tau}(n, k)$ of arcs in a directed k-failure-safe total communication network is given by $\vec{\tau}(n, k) = (k+2)n-2$. The problem of finding $\vec{\tau}(n, 0)$ is called the telegraph problem. See [4].

2. Communication to a single vertex. In this section, we prove the following theorem.

THEOREM 2.1. *Let $\mu(n, k)$ be the minimum size of a communication network on n vertices in which every vertex communicates k-failure-safe with a given vertex. Then*

$$\mu(n,k)=\begin{cases}\left\lceil\left(\dfrac{k+2}{2}\right)(n-1)\right\rceil, & k\leqq n-2,\\[2ex] \left\lceil\left(\dfrac{k+1}{2}\right)n\right\rceil, & k\geqq n-2.\end{cases}$$

Proof. An *ascending tree rooted at v* is a tree whose edges are ordered such that there is an ascending path from every vertex in the tree to v. For $k=0$, an ascending tree on n vertices clearly gives the minimum solution. For $k>0$ there are two cases: $k\leqq n-2$ and $k \geqq n-2$.

Case 1. $k \leqq n-2$. We first show by construction that $\mu(n, k) \leqq \lceil((k+2)/2)\cdot(n-1)\rceil$. Consider a multigraph G with vertices $v_1, v_2, \cdots, v_n$ which is the edge disjoint union of graphs T and T' where T is a spanning tree such that every vertex of G different from v_1 is joined to v_1 and T' is a simple graph having degree k at every vertex different from v_1 and degree either 0 or 1 at v_1. Graph T' can be constructed as follows: For *k even* v_iv_j is an edge of T' precisely when $i\neq 1, j \neq 1$ and $i-j$ is congruent mod $(n-1)$ to an element of $\{1, -1, 2, -2, \cdots, k/2, -k/2\}$. In the case *$k$ odd, n odd,* v_iv_j is an edge of T' precisely when $i \neq 1, j\neq 1$ and $i-j$ is congruent mod $(n-1)$ to an element of $\{1, -1, 2, -2, \cdots, (k-1)/2, -(k-1)/2, (n-1)/2\}$. Finally in the case *$k$ odd, n even* the pairs $v_2v_{n/2+1}, v_3v_{n/2+2}, \cdots, v_{n/2}v_{n-1}$ and v_1v_n are edges of T'. Furthermore, if $i\neq 1, j\neq 1$ and $i-j$ is congruent mod $(n-1)$ to an element of $\{1, -1, 2, -2, \cdots, (k-1)/2, -(k-1)/2\}$ then v_iv_j is an edge of T'.

We now order the edges of G by first ordering the edges of T' arbitrarily and then ordering the edges of T arbitrarily. With this ordering G is a communication network in which every vertex communicates k-failure-safe with vertex v_1. To prove this it is sufficient to show that there are $k+1$ pairwise edge disjoint ascending paths from any vertex v_i to v_1. Let $v_{i_1}, v_{i_2}, \cdots, v_{i_k}$ be the k vertices adjacent to v_i in T'. The $k+1$ paths $v_iv_1, v_iv_{i_1}v_1, v_iv_{i_2}v_1, \cdots, v_iv_{i_k}v_1$ are pairwise edge disjoint ascending paths from v_i to v_1.

Since the tree T has $n-1$ edges and the graph T' has by a simple degree counting argument $\lceil k(n-1)/2\rceil$ edges it follows that G has $\lceil(k+2)(n-1)/2\rceil$ edges. Hence $\mu(n, k) \leqq \lceil((k+2)/2)(n-1)\rceil$.

We now show that $\mu(n, k)\geqq \lceil((k+2)/2)(n-1)\rceil$. Let G be any communication network in which every vertex communicates k-failure-safe with a single vertex v_1. Let v be any vertex different from v_1 and let vw be the edge of highest order which is incident with v. If P is any ascending path which terminates at v_1 and if P contains the edge vw, then in P vertex w must lie between v and v_1. Thus replacing vw with vv_1 preserves the property that every vertex communicates k-failure-safe with v_1. Repeated replacements of this kind results in a graph G' where every vertex other than v_1 is adjacent to v_1. Clearly every vertex of G' different from v_1 must have degree at least $k+1$. It follows that the sum of the degrees of the vertices in G' is at least $(n-1)+(n-1)(k+1)$. Therefore the number of edges of G' is at least $\lceil (n-1)(k+2)/2 \rceil$. This shows that the number of edges of G is at least $\lceil (n-1)(k+2)/2 \rceil$.

Case 2. $k \geqq n-2$. The following upper bound construction shows that $\mu(n, k) \leqq \lceil ((k+1)/2)n \rceil$. Let λ, r be positive integers such that $k+1 = \lambda(n-1)+r$, $r < n-1$. Let R be a simple graph which has degree r at every vertex different from v_1 and degree either r or $r+1$ at v_1. Let K_n be the complete graph on n vertices. Consider the multigraph G on vertex set V which is the union of λ copies of K_n and one copy of R, i.e., two vertices u and w are joined with $\lambda + l$ multiple edges where l is the number of edges joining vertices u and w in R ($l = 0$ when there are no such edges). We now make G a communication network by imposing an order on its edges. First, in an arbitrary fashion, order all the edges of R not incident with vertex v_1. Then order all the edges of the complete graphs not incident with vertex v_1 in an arbitrary way. Finally order all the edges incident with v_1 in an arbitrary way. The reader can easily verify that there exists $k+1$ pairwise edge disjoint ascending paths (of lengths 1, 2 and 3) from any vertex to v_1. Thus every vertex of G communicates k-failure-safe with v_1. Since every vertex in G has degree $k+1$ except v_1 which has degree either $k+1$ or $k+2$ it follows that the number of edges of G is $\lceil ((k+1)/2)n \rceil$. Hence, $\mu(n, k) \geqq \lceil ((k+1)/2)n \rceil$.

Conversely let G be a graph in which every vertex communicates k-failure-safe with v_1. It is clear that every vertex must have degree at least $k+1$. Hence G must have at least $\lceil ((k+1)/2)n \rceil$ edges and thus $\mu(n, k) \geqq \lceil ((k+1)/2)n \rceil$.

This completes the proof of Theorem 2.1.

3. *k*-failure-safe total communication networks.

In this section we consider communication networks where every vertex communicates k-failure-safe with every other vertex. Let $\tau(n, k)$ denote the minimum size of such a network. As mentioned in the introduction, the problem of finding $\tau(n, 0)$ is the telephone problem and it is well known that $\tau(n, 0) = 2n-4$. The following theorem gives upper and lower bounds for $\tau(n, k)$ when $k \geqq 1$.

THEOREM 3.1. *Let $\tau(n, k)$ be the minimum size of a k-failure-safe total communication network where $k \geqq 1$. Then $\tau(n, k)$ satisfies*

$$\left\lceil \left(\frac{k+4}{2}\right)(n-1) \right\rceil - 2\lceil \sqrt{n} \rceil + 1 \leqq \tau(n, k) \leqq \left\lfloor \left(k+\frac{3}{2}\right)(n-1) \right\rfloor, \quad k \leqq n-2,$$

$$\left\lceil \left(\frac{k+3}{2}\right)n \right\rceil - 2\lceil \sqrt{n} \rceil \leqq \tau(n, k) \leqq \left\lfloor \left(k+\frac{3}{2}\right)(n-1) \right\rfloor, \quad k \leqq n-2.$$

We first show that the upper bound holds by construction. For n odd, we construct a communication network G as follows. In the case when i is odd, $i \neq 1$, join vertex v_i with k edges to vertex v_1 labeling them $3, 5, \cdots, 2k+1$. In the case when i is even join v_i with $k+1$ edges to v_1 labeling them $2, 4, 6, \cdots, 2k+2$. Further join v_i and v_{i+1} with two edges, one edge labeled 1 and the other edge labeled $2k+3$ for $i = 2, 4, \cdots, n-1$. The labels determine a partition of the edges into $2k+3$ classes. Edges in the same class are ordered arbitrarily but for $i < j$ all the edges in class i are ordered before any of the edges in class j.

Consider any two distinct vertices v_i and v_j in communication network G. It is an easy exercise for the reader to verify that there exist $k+1$ pairwise edge disjoint ascending paths from v_i to v_j. If n is even then we construct G as follows. First do the above construction for $n-1$ vertices. Then join vertex v_n to v_1 with $k+1$ edges, one from each of the $k+1$ new classes $2^+, 3^+, 5^+, 7^+, \cdots, (2k+1)^+$, where an edge in class j^+ is ordered so that it is larger than any edge in class j and smaller than any edge in class $j+1$. Again it is easy to verify that every vertex communicates k-failure-safe with every other vertex. The number of edges in G is $\lfloor (k+\frac{3}{2})(n-1) \rfloor$. This gives the upper bound for $\tau(n, k)$.

We now obtain the lower bound. Let G be a communication network in which every vertex communicates k-failure-safe with every other vertex. Let $E(r)$ be the set of the first r edges in the linear order. Let $E'(r)$ denote the edges that follow these. Set $\mu = \mu(n, k)$. For a positive integer ξ, $\xi \leqq \mu$, let H be subgraph whose vertex set includes all the vertices of G and whose edge set is $E'(\mu - \xi)$. Let K be the connected component of H with the fewest number of edges and let l be the number of edges of K. Let C be the subgraph of G whose vertex set includes all the vertices of G and whose edges consist of the set $E(\mu - \xi)$ together with the l edges of K. (Subgraph C may contain isolated vertices.) Let v be any vertex of K. It is clear that in C every vertex communicates k-failure-safe with v since every vertex communicates k-failure-safe with v in G. Hence C has at least μ edges. But the number of edges of C is $\mu - \xi + l$. Therefore $l \geqq \xi$. Since K was the smallest component of H subgraph H has at least $n - n/\xi$ edges. (It is an easy exercise to verify that a graph has at least $n - n/\xi$ edges if each of its components has size at least ξ.) This implies that G has at least $\mu + n - \xi - n/\xi$ edges. To optimize choose $\xi = \lceil \sqrt{n} \rceil$. Then we have that $|E(G)| \geqq \mu + n - 2\lceil \sqrt{n} \rceil$, i.e., $|E(G)| \geqq \mu(n, k) + n - 2\lceil \sqrt{n} \rceil$. By employing the formula for $\mu(n, k)$ we obtain the lower bound for Theorem 3.1.

For $k = 1$ Theorem 3.1 gives $\tau(n, k)$ to within $2\sqrt{n}$. For $k > 1$ the bounds of Theorem 3.1 become increasingly less tight as k increases. We propose the following conjecture.

Conjecture. For $k \geqq 1$

$$\tau(n, k) = (k+\tfrac{3}{2})n - c$$

where c is bounded as n goes to infinity.

4. Directed communication networks. In this section we prove two theorems on directed communication networks which are the analogs of Theorems 2.1 and 3.1.

THEOREM 4.1. *Let $\vec{\mu}(n, k)$ be the minimum size of a directed communication network on n vertices in which every vertex v communicates k-failure-safe with a single vertex v_1. Then*

$$\vec{\mu}(n, k) = (k+1)(n-1).$$

Proof. Let D be the multidigraph on vertices $v_1, v_2, \cdots, v_n$ such that vertex v_i is joined to vertex v_1 with $k+1$ arcs directed toward v_1, $i = 2, 3, \cdots, n$. Clearly every vertex communicates k-failure-safe with every other vertex and D has $(k+1)(n-1)$ arcs. This shows that $\vec{\mu}(n, k) \leqq (k+1)(n-1)$.

Conversely, let D be a directed communication network such that every vertex communicates k-failure-safe with every other vertex. Clearly, D must have out-degree at least $k+1$ at every vertex v_i, $i \neq 1$, and thus D must have at least $(k+1)(n-1)$ arcs. This shows that $\vec{\mu}(n, k) \geqq (k+1)(n-1)$.

THEOREM 4.2. *Let $\vec{\tau}(n, k)$ be the minimum size of a directed k-failure-safe total communication network. Then*

$$\vec{\tau}(n, k) = (k+2)n - 2.$$

Proof. We first show that $\vec{\tau}(n, k) \leqq (k+2)n - 2$ by construction. Let $v_1, v_2, \cdots, v_n$ denote the n vertices. We construct a directed communication network D as follows:

Join vertex v_i to vertex v_{i+1} with $k+2$ arcs directed toward v_{i+1} for $i=1,2,\cdots,n-2$. Join v_{n-1} to v_n with $k+1$ arcs directed toward v_n and join v_n to v_1 with $k+1$ arcs directed toward v_1. We now order the arcs of D. The $k+2$ arcs joining vertex v_i to vertex v_{i+1} are ordered $i, n+i, 2n+i, \cdots, (k+2)n+i$ for $i=1,2,\cdots,n-2$. The $k+1$ arcs joining vertex v_{n-1} to vertex v_n are ordered $n-1, 2n-1, \cdots, (k+1)n-1$ and the $k+1$ arcs joining vertex v_n to vertex v_1 are ordered $n, 2n, \cdots, (k+1)n$. We leave it to the reader to verify that there are $k+1$ pairwise edge disjoint ascending directed paths from any vertex of D to any other vertex of D, i.e., D is a directed k-failure-safe total communication network. Since D has $(k+2)n-2$ edges $\vec{\tau}(n,k) \leqq (k+2)n-2$.

We now show that $\vec{\tau}(n,k) \geqq (k+2)n-2$. Let D be a directed k-failure-safe total communication network. Consider the $n-2$ arcs which occur first in the linear order and let A be the subdigraph consisting of these $n-2$ arcs. Since a connected graph must have at least $n-1$ edges A is not connected. Therefore for any vertex v of D there is some vertex v' which does not communicate with v using only the arcs in A. Since D is a directed k-failure-safe total communication network, and since the arcs from A occur first in the linear order it follows that there are at least $k+1$ arcs directed toward v which do not belong to A for every vertex v of D. This implies that there are at least $(k+1)n$ arcs not lying in A. Hence D has at least $(n-2)+(k+1)n = (k+2)n-2$ arcs showing that $\vec{\tau}(n,k) \geqq (k+2)n-2$.

This proves Theorem 4.2.

Acknowledgments. We thank the referee for several valuable comments. We thank Daniel Kleitman for many illuminating discussions.

REFERENCES

[1] B. Baker and R. Shostah, *Gossips and telephones*, Discrete Math., 2 (1972), pp. 191-193.

[2] R. T. Bumby *A problem with telephones*, this Journal, 2 (1981), pp. 13-19.

[3] A. Hajnal, E. C. Milner and E. Szemeredi, *A cure for the telephone disease*, Canad. Math. Bull., 15 (1976), pp. 447-450.

[4] F. Harary and A. J. Schwenk, *The communication problem on graphs and digraphs*, J. Franklin Inst., 297 (1974), pp. 491-495

[5] D. J. Kleitman and J. B. Shearer, *Further gossip problems*, Discrete Math., 30 (1980), pp. 151-156.

[6] R. Tidjeman, *On a telephone problem*, Nieuw Arch. Wisk., 3 (1971), pp. 188-192.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

003

ADJACENCY MATRICES*

COLIN D. WALTER†

Abstract. For a graph Γ with vertex set V an algebra of adjacency matrices is defined and viewed as an equivalence relation on $V \times V$ with certain nice properties. This can be used in algorithms to find automorphisms of graphs and isomorphisms between graphs. It also provides intersection numbers independent of the labelling on V which determine the similarity class of the adjacency algebra.

AMS(MOS) subject classification. 05C50

Introduction. This article has two main objectives. The first is to associate as high a dimensional algebra of $n \times n$ matrices as possible with the adjacency matrix of a labelled graph Γ on n vertices. In this way a set of intersection numbers is obtained which is an invariant for the isomorphism class of Γ. The other aim is to show that these intersection numbers provide a finer decomposition into equivalence classes of graphs than do graph spectra, even with the more general definition given here. It therefore seems likely that nice classification theorems must exist using these numbers, giving more powerful results than from spectra. Indeed the theory of distance transitive graphs illustrates this (see [1]). However, such results are not given here. What is provided is the step from a given graph to a coherent configuration as defined by D. G. Higman [4] and one can then apply his theory. He gives some applications.

The associated algorithm which tests for isomorphism by computing these numbers (implicitly) has order at worst $n^3 \log n$ and can be applied recursively to the subgraphs obtained by deleting vertices until isomorphism is established or confuted. The calculation is then producing generalised intersection numbers corresponding successively to ordered pairs, triples, quadruples, etc., of vertices. This points to the correct generalisation to yield invariants which completely determine the isomorphism class of the graph.

Sections 1 to 3 are definitions and elementary properties. Section 4 starts with a couple of well-known results which can be traced back to Frobenius [3]. From them is deduced that intersection numbers are more discriminating than the spectrum. In § 5 these numbers are shown to be equivalent to knowledge of the regular representation, for which a symmetric definition and an easy method of computation are given. Lastly, in § 6, the names of the labels, hitherto ignored, are traced to ensure that an isomorphism preserves not just the equivalence classes of edges carrying the same label, but also the label itself.

The starting point of this paper was a talk by Charles R. Johnson on a joint work of his with Morris Newman [5]. The author would especially like to thank T. J. Laffey for many helpful conversations during its development.

The intersection numbers are obtained in the following way. Let A be an adjacency matrix of a graph Γ. Any automorphism of Γ acts as a similarity transformation by a permutation matrix on A. Thus such transformations act trivially on the algebra generated by all such A for the given graph. A generic matrix of this algebra can be used to partition the vertex set V of Γ into subsets $V_1, V_2, \cdots, V_t$ with the property that any automorphism of Γ restricted to V_i maps onto V_i. The V_i are unions of orbits under the automorphism group.

This can be expressed abstractly using equivalence relations on $V \times V$: giving a "colour" to each edge and vertex. There is a smallest refinement of this colouring of

* Received by the editors August 2, 1983, and in revised form August 6, 1984.

† Department of Mathematics, University College, Belfield, Dublin 4, Ireland.

$V \times V$ with a property corresponding to closure under multiplication of matrices. This is called here the *completion* of the colouring, but is just a *coherent configuration* in Higman's terminology.

The formulae in terms of colours for the product of two matrices in this algebra define the intersection numbers and determine the algebra up to similarity. Thus they are identical for isomorphic graphs and can be used as a test for isomorphism. The adjacency algebra defined in this way is larger than the usual one, being generated by all possible adjacency matrices instead of a single 0, 1-matrix. It is big enough to show how closely connected are the ideas of similarity, co-spectrality, and intersection numbers.

Addendum. The author would like to note that associating a coherent configuration with a graph is the subject of [10]. This does not seem to be well known despite its reference in [11]. The first few sections here describe the method.

1. Colourings.

DEFINITION 1.1. Let V be a finite set and c an equivalence relation on $V \times V$ with r equivalence classes. Then c is called an *r-colouring* or *colouring* of V and the equivalence classes are called the *colours* of c. The set of such classes will be denoted by c and the class of $(i, j) \in V \times V$ by $c(i, j)$. This should be distinguished from $c((i, j))$, also called the *colour* of (i, j), which is always the image of $c(i, j)$ under an injective map. Elements of V are identified with the diagonal of $V \times V$ and called *vertices*, whilst off-diagonal elements are called *edges.*

For example, let Γ be a graph on V with edge set E. Then Γ yields a 3-colouring of V whose colours are V, E, and $(V \times V)\backslash(E \cup V)$.

DEFINITION 1.2. For $n = |V|$ and a commutative ring R containing the integers $\mathbb{Z}$, let $M_V(R)$ be the set of $n \times n$ matrices with entries in R whose rows and columns are indexed by V. Any injective map $c \to R$ with $c(i, j) \mapsto c((i, j))$ defines a matrix $A = (a_{ij}) \in M_V(R)$ by $a_{ij} = c((i, j))$. Such a matrix is called an *adjacency matrix* of c. Conversely, given a matrix $A = (a_{ij}) \in M_V(R)$ there is a uniquely determined colouring for which A is an adjacency matrix, namely that given by $c(i, j) = c(k, l) \Leftrightarrow a_{ij} = a_{kl}$ for all $i, j, k, l \in V$. The colouring so obtained is denoted c_A. If the set of distinct entries of A are algebraically independent over $\mathbb{Z}$ (as a subring of R) then A is called a *generic matrix* of the colouring it defines. A set of generic matrices (not necessarily for the same colouring) are called *independent* if the entries in each matrix are distinct from the entries in every other matrix and the set of distinct entries from all the matrices is algebraically independent over $\mathbb{Z}$.

LEMMA 1.3. *Let c, d be colourings of V. Then $c = d$ if and only if, $c((i, j)) = c((k, l)) \Leftrightarrow d((i, j)) = d((k, l))$ for all $i, j, k, l \in V$.*

Example 1.4. The colourings with adjacency matrices

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{pmatrix} \quad \text{and} \quad A^T = \begin{pmatrix} 1 & 3 & 2 \\ 2 & 1 & 3 \\ 3 & 2 & 1 \end{pmatrix}$$

are equal.

DEFINITION 1.5. (i) There is a *partial ordering* $\leqq$ of colourings given by

$$c \leqq d \text{ if, and only if, } c(i, j) \supseteq d(i, j) \text{ for all } (i, j) \in V \times V.$$

(ii) The *sum* or *join* $c + d$ is defined by

$$(c + d)(i, j) = c(i, j) \cap d(i, j)$$

and the *meet* $c \wedge d$ is defined so that $(c \wedge d)(i, j)$ is the smallest union of colours of c containing (i, j) which is also a union of colours of d.

(iii) The rank of a colouring c is the number $|c|$ of equivalence classes of c. Clearly, $1 \leqq |c| \leqq n^2$ for $n = |V|$.

LEMMA 1.6. (i) *The colourings of V form a lattice under $\leqq$ with meet and join as above.*

(ii) *$c + d$ is the least upper bound for c and d. In particular, $c \leqq c + d$ and $d \leqq c + d$. Also $c + d$ is the colouring defined by the sum of independent generic matrices for c and d. Moreover, $c + d = c$ if $d \leqq c$.*

(iii) *$c \wedge d$ is the greatest lower bound for c and d. In particular, $c \wedge d \leqq c$ and $c \wedge d \leqq d$ with $c \wedge d = c$ if $c \leqq d$.*

(iv) *The map $c \mapsto |c|$ is order preserving, i.e. $c < d$ implies $|c| < |d|$. Also $|c \wedge d| \leqq |c| \leqq |c + d| \leqq |c|\,|d|$.*

DEFINITION 1.7. (i) There is a unique minimal colouring c_0 corresponding to the zero matrix. This is a 1-colouring with $c_0(i, j) = V \times V$.

(ii) There is a unique maximal colouring c_V which is defined by $c_V(i, j) = \{(i, j)\}$. It has $|V|^2$ colours.

(iii) The *identity colouring* c_I is that which corresponds to the identity matrix. It is a 2-colouring with $c_I(i, i) = V \subseteq V \times V$ and $c_I(i, j) = V \times V \setminus V$ for $i \neq j$.

DEFINITION 1.8. The *transpose colouring* c^T is defined by $c^T(i, j) = c(j, i)^T$ where $S^T = \{(i, j) | (j, i) \in S\}$ for any subset S of $V \times V$. A colouring c is called *symmetric* if $c = c^T$ and *totally symmetric* if $c(i, j)^T = c(i, j)$ for all $i, j \in V$. Because $(j, i) \in c(i, j)^T$, c is totally symmetric precisely when $c(i, j) = c(j, i)$ for all $i, j \in V$.

Remark 1.9. Suppose A is a generic matrix for c. Then A^T is a generic matrix for c^T.

Example 1.4 illustrates a symmetric colouring which does *not* arise from a symmetric matrix. The totally symmetric colourings are characterised by having symmetric adjacency marices, whilst the symmetric colourings are characterised by having their set of adjacency matrices closed under the transpose mapping.

The product cd of two colourings is defined as that obtained from the product of independent generic matrices for c and d. Hence we have the following definition.

DEFINITION 1.10. The *product* cd of two colourings c, d of V is defined by its injective image

$$cd((i, j)) = \{c(i, t) \times d(t, j) | t \in V\}$$

or, equivalently,

$$cd((i, j)) = \{(c((i, t)), d((t, j))) | t \in V\}$$

where the elements are counted with appropriate multiplicity. *All such sets from here on will be assumed to have multiplicities attached to their elements, i.e. they are* *__multisets__* *or* *__bags__*.

In computations as in Example 1.4 the values $c((i, j))$ are usually integers. Then the product class $cd(i, j)$ consists of those directed edges (i, j) yielding the same $|V|$-tuple of pairs $(c(i, t), d(t, j))$ sorted into order. Thus, if c, d are the colourings in Example 1.4, then $cd(2, 3)$ is the set of edges giving the triple $(13, 21, 32)$. If generic matrices are used, one has $x_3y_2 + x_1y_3 + x_2y_1$ representing this class. Ordering the terms lexicographically and recording only subscripts yields the previous triple.

THEOREM 1.11. (i) *The sum and product operations satisfy the usual associative and distributive axioms of rings. Addition is commutative but multiplication is not commutative if $|V| > 1$.*

(ii) *$cc_I \geqq c$ and $c_Ic \geqq c$.*

(iii) *$ce \leqq df$ if $c \leqq d$ and $e \leqq f$ for colourings c, d, e, f.*

(iv) *$c+d \leqq cd$ if $c \geqq c_I$ and $d \geqq c_I$.*

Proof. (i) Generic matrices which determine colourings satisfy the named axioms of ring theory. Hence the colourings themselves satisfy these axioms. For $|V|>1$ let c be the colouring with generic matrix $A=(a_{ij})$ such that $a_{1j}=x$ and $a_{ij}=y$ for $i \neq 1$. Easily $cc^T \neq c^Tc$ since the former is a 4-colouring and the latter the 1-colouring.

(ii) $(k, l) \in cc_I(i, j)$ implies

$$\{c(k, t) \times c_I(t, l) | t \in V\} = \{c(i, t) \times c_I(t, j) | t \in V\}.$$

Equating terms which contain the diagonal $V=c_I(t, t)$ gives $c(k, l)=c(i, j)$ and hence $(k, l) \in c(i, j)$. Thus $cc_I(i, j) \subseteq c(i, j)$ and $cc_I \geqq c$. By symmetry $c_Ic \geqq c$.

(iii) $(k, l) \in df(i, j)$ implies $\{d(k, t) \times f(t, l) | t \in V\} = \{d(i, t) \times f(t, j) | t \in V\}$ and hence $\{c(k, t) \times e(t, l) | t \in V\} = \{c(i, t) \times e(t, j) | t \in V\}$. Therefore $(k, l) \in ce(i, j)$. Thus $df(i, j) \subseteq ce(i, j)$ and $ce \leqq df$, as required.

(iv) From (ii) and (iii) $c \leqq cc_I \leqq cd$ and $d \leqq c_Id \leqq cd$, giving $c+d \leqq cd$.

LEMMA 1.12. (i) *$c^{TT}=c$;*

(ii) *$(cd)^T=d^Tc^T$ and $(c+d)^T=c^T+d^T$;*

(iii) *cc^T and $c+c^T$ are symmetric;*

(iv) *$c \leqq d$ implies $c^T \leqq d^T$ and vice versa;*

(v) $|c|=|c^T|$.

PROPOSITION 1.13. *Let c^r be the product of c with itself r times for $r \in \mathbb{Z}$, $r>0$, and set $c^0=c_I$. Take $n=|V|>1$.*

(i) *If $c \geqq c_I$ then there is a positive integer $m<n^2$ such that $c^m=c^{m+r}$ for all $r \geqq 0$.*

(ii) *For each colouring c there are positive integers m, p bounded by functions of n such that $c^r=c^{p+r}$ for all $r \geqq m$.*

Proof. (i) By Theorem 1.11(ii) and (iii), $c^r \leqq c^rc_I \leqq c^{r+1}$ for all $r \geqq 0$. If $c_I=c^0<c^1<\cdots<c^r$ then $2=|c^0|<|c^1|<\cdots<|c^r|$ by Lemma 1.6(iv) and so $|c^r| \geqq r+1$. Now $|c^r| \leqq n^2$ yields $r<n^2$. Hence there is a maximal value $r=m$ with this property, i.e. $c^m=c^{m+1}$, which gives $c^m=c^{m+r}$ for all $r \geqq 0$.

(ii) This is automatic from the finitude of the number of colourings for fixed n.

DEFINITION 1.14. In Proposition 1.13 the minimal m satisfying (i) is called the *order* of c, and the minimal value of p satisfying (ii) is called the *period* of c.

The *completion* of c is $\bar{c}=(c+c^T+c_I)^{n^2}$ for $n=|V|$, and c is called *complete* if $c=\bar{c}$.

Remarks 1.15. Note that $c+c^T+c_I \geqq c_I$. Thus, by Proposition 1.13, its period is 1 and $\bar{c}=(c+c^T+c_I)^m$ where m is the order of $c+c^T+c_I$. In computations $\bar{c}$ is obtained by successively squaring $c+c^T+c_I$. The rth squaring gives $(c+c^T+c_I)^{2^r}$ and so $\bar{c}$ results after at most $\log_2(n^2-1)$ steps. The computation terminates when squaring returns the same colouring.

$\bar{c}$ is the maximal colouring obtainable from c using c_I and the operations so far defined because of the next theorem.

THEOREM 1.16. (i) *$\bar{c}^2=\bar{c}$; $\bar{c}+\bar{c}=\bar{c}$; and $\bar{c}^T=\bar{c}$,*

(ii) *If $c_1 \leqq \bar{c}$, $c_2 \leqq \bar{c}$ then $c_1c_2 \leqq \bar{c}$, $c_1+c_2 \leqq \bar{c}$ and $c_1^T \leqq \bar{c}$,*

(iii) *If $c \leqq d$ then $\bar{c} \leqq \bar{d}$,*

(iv) *$\bar{\bar{c}}=\bar{c}$,*

(v) *$\bar{c}c_I=\bar{c}=c_I\bar{c}$ and $\bar{c} \geqq c_I$,*

(vi) *c is complete if, and only if, $c \geqq c_I$, $c^T=c$ and $c^2=c$.*

THEOREM 1.17. *Suppose c is the totally symmetric 2- or 3-colouring of a regular graph with adjacency matrix A and c_i is the colouring associated with A^i. Then, for $n=|V|$, $\bar{c}=c_0+c_1+\cdots+c_{n-1}$.*

Proof. By Theorem 1.16, $c_i \leqq \bar{c}$ and therefore $c_p \leqq \bar{c}$ where $c_p = c_0 + c_1 + \cdots + c_{n-1}$. Now c has generic matrix $xI + yJ + zA$ with $JA = AJ = dJ$ for some $d \in \mathbb{Z}$ and $J^2 = nJ$. So any polynomial in I, J, A is a linear combination of $A^0, A^1, \cdots, A^{n-1}$ and J by the Cayley-Hamilton theorem. Since $\bar{c} = (c + c_I)^i$ for i large enough, $\bar{c}$ has an adjacency matrix of this form and $\bar{c} \leqq c_p$, giving $\bar{c} = c_p$.

Remark 1.18. Complete colourings are the same as coherent configurations in the sense of D. G. Higman [4]. The intersection numbers he has are just the multiplicities of the various terms in each entry of a product of two independent generic matrices. Thus completion provides a natural and easy way of associating a coherent configuration with any graph. The completion $\bar{c}$ is the minimal coherent configuration which is a refinement of c. If $\bar{c}$ is totally symmetric then it is an association scheme in the sense of Bose and Shimamoto [2]. If c is obtained from a strongly regular graph, then $\bar{c} = c$ (see J. J. Seidel [8]).

2. Automorphisms.

DEFINITION 2.1. Let S_V denote the group of permutations of V. S_V acts naturally on $V \times V$ by $\sigma(i,j) = (\sigma i, \sigma j)$. Thus ·$\sigma T$ is well-defined for subsets T of $V \times V$ and $\sigma \in S_V$. In particular, a colouring c with classes $c_1, c_2, \cdots, c_r$ yields a colouring σc with classes $\sigma c_1, \sigma c_2, \cdots, \sigma c_r$ where $\sigma c_k = \{(\sigma i, \sigma j) | (i,j) \in c_K\}$. The *(strict) automorphism group* Aut*c of a colouring c is the subgroup of S_V consisting of permutations which leave the colours fixed, i.e.,

$$\sigma \in \text{Aut}^*\, c \quad \Leftrightarrow \quad \sigma c_i = c_i \quad \text{for each colour } c_i \text{ of } c.$$

Of less interest here is the group Aut $c = \{\sigma \in S_V | \sigma c = c\}$ which may include automorphisms which permute the colours nontrivially. For a matrix $A = (a_{ij})$ with associated colouring c, $\sigma A = (\sigma a_{ij})$ is associated with σc and so has entries $\sigma a_{ij} = a_{\sigma^{-1}i, \sigma^{-1}j}$. Then, obviously, Aut* $c = \{\sigma \in S_V | \sigma A = A\}$ whilst Aut c consists of those σ for which σA is also an adjacency matrix of c.

LEMMA 2.2. (i) $\sigma c + \sigma d = \sigma(c + d)$; $(\sigma c)(\sigma d) = \sigma(cd)$; $\sigma(c^T) = (\sigma c)^T$ *for* $\sigma \in S_V$;

(ii) Aut* $(c + d) =$ Aut* $c \cap$ Aut*d;

(iii) Aut* $(cd) =$ Aut* $c \cap$ Aut* d *if* $c \geqq c_I$ *and* $d \geqq c_I$;

(iv) Aut* $(c^T) =$ Aut* c;

(v) $c \leqq d$ *implies* Aut* $c \supseteq$ Aut* d.

Proof. (i), iv) and (v) are clear.

(ii) If A, B are independent generic matrices for c, d then

$$\sigma \in \text{Aut}^*\,(c+d) \Leftrightarrow \sigma(A+B) = A + B \Leftrightarrow (\sigma A = A \text{ and } \sigma B = B) \Leftrightarrow \sigma \in \text{Aut}^*\, c \cap \text{Aut}^* d.$$

(iii) Here $\sigma \in$ Aut* $c \cap$ Aut* d implies $\sigma(AB) = (\sigma A)(\sigma B) = AB$ and so $\sigma \in$ Aut* cd. Thus, Aut* $c \cap$ Aut* $d \subseteq$ Aut* cd without restriction. Assuming (v) and using (ii) with Theorem 1.11(iv) gives Aut* $c \cap$ Aut* $d =$ Aut* $(c + d) \supseteq$ Aut* cd and so equality must hold.

THEOREM 2.3. Aut* $c =$ Aut* $(c + c^T + c_I) =$ Aut* $\bar{c}$.

3. Complete colourings.

LEMMA 3.1. *Suppose c is complete.*

(i) $c(i,j) \neq c(k,l)$ *if* $\delta_{ij} \neq \delta_{kl}$ (*Kronecker delta*).

(ii) *If* $c(i,j) = c(k,l)$ *then there is a permutation* $\sigma \in S_V$ *with* $c(i,t) = c(k, \sigma t)$ *and* $c(t,j) = c(\sigma t, l)$ *for all* $t \in V$.

(iii) *If* $c(i,j) = c(k,l)$ *then* $c(i,i) = c(k,k)$ *and* $c(j,j) = c(l,l)$.

Proof. (i) is immediate from $c \geqq c_I$. Using the definition of product and $c^2 = c$ gives $\{c(i,t) \times c(t,j) | t \in V\} = c^2((i,j)) = c^2((k,l)) = \{c(k,t) \times c(t,l) | t \in V\}$. Any bijection between these two bags which preserves colours determines a suitable $\sigma \in S_V$ in (ii). In particular, restricting σ to diagonal classes yields (iii).

THEOREM 3.2 [3, § 2.10]. *If $V = V_1 \dot{\cup} V_2 \dot{\cup} \cdots \dot{\cup} V_t$ is the partition of V induced by the diagonal classes of a complete colouring c then each block $V_i \times V_j$ is a union of colours of c.*

COROLLARY 3.3. *With the hypotheses and notation of Theorem 3.2, the permutation $\sigma \in S_V$ in Lemma 3.1(ii) satisfies $\sigma V_i = V_i$ for each i.*

COROLLARY 3.4. *Suppose V_1 and V_2 are diagonal classes (possibly equal) for a complete colouring c. Then $\{c(i, t) | t \in V_2\}$ and $\{c(t, j) | t \in V_1\}$ are independent of $i \in V_1$ and $j \in V_2$ respectively. The multiplicities of a colour c_k in $i \times V_2$ and $V_1 \times j$ are related by*

$$|c_k \cap (i \times V_2)| \, |V_1| = |c_k \cap (V_1 \times j)| \, |V_2|.$$

If $c_k \subseteq V_1 \times V_2$ then $|V_1|$ and $|V_2|$ divide $|c_k|$.

Proof. For $i, i' \in V_1$, $c(i, i) = c(i', i')$. So, by Lemma 3.1, there is a $\sigma \in S_V$ with $c(i, t) = c(i', \sigma t)$ for all $t \in V$. By Corollary 3.3, σ restricts to $\sigma_2 : V_2 \to V_2$. Hence $\{c(i, t) | t \in V_2\}$ is independent of $i \in V_1$. Independence for the second set follows similarly or by applying the transpose. This immediately gives the equation relating multiplicities, both sides having cardinality $|c_k \cap V_1 \times V_2|$. The last part is now clear.

THEOREM 3.5. *The restriction c_i of a complete colouring c to $V_i \times V_i$ for a diagonal class V_i of c is a complete colouring with one diagonal class.*

Proof. Clearly $c_i \geqq c_I$ and $c_i = c_i^T$ because these properties hold for c. Suppose $c_i(j, k) = c_i(r, s)$. Then $c(j, k) = c(r, s)$ and by Corollary 3.3 the permutation $\sigma \in S_V$ defined in Lemma 3.1 restricts to a map $\sigma_i : V_i \to V_i$ such that $c(j, t) = c(r, \sigma_i t)$ and $c(t, k) = c(\sigma_i t, s)$ for $t \in V_i$. So $c_i^2((j, k)) = \{c(j, t) \times c(t, k) | t \in V_i\} = \{c(r, t) \times c(t, s) | t \in V_i\} = c_i^2((r, s))$. This means $c_i^2 \leqq c_i$ and hence $c_i = c_i^2$. Thus c_i is complete.

Remark 3.6 [3, § 8]. In the same way a complete colouring restricts to a complete colouring on any union of its diagonal classes.

DEFINITION 3.7. The number of colours on the diagonal of a colouring $c \geqq c_I$ is denoted $\|c\|$. A complete colouring is called *regular* if $\|c\| = 1$ ("homogeneous" in the terminology of Higman).

Remark 3.8. $\|c\|^2 \leqq |c|$ for complete colourings.

4. Adjacency algebras and determinants. If we regard a matrix in $M_V(R)$ as a map $V \times V \to R$ in the obvious way, then the adjacency matrices of a colouring c are the maps $\phi : V \times V \to R$ for which every $\phi^{-1}(r)$, $r \in R$, is either the empty set or a colour of c. The adjacency matrices for all colourings $d \leqq c$ are the maps $\phi : V \times V \to R$ which are constant on each colour of c, that is, $\phi^{-1}(r)$ is a union of colours of c for all $r \in R$. Such matrices form a free R-module $M_c = M_c(R)$ of rank $|c|$. Certainly $I \in M_c$ if, and only if, $c \geqq c_I$. Indeed, $c \leqq d$ if, and only if, $M_c \subseteq M_d$. The most important observation is that *M_c is a ring if c is complete.* When R is a field and c is complete M_c is therefore an algebra. $M_{\bar{c}}$ is the *adjacency ring* (or algebra) over R of the colouring c.

THEOREM 4.1 (see e.g. Higman [4]). *For a subfield K of the complex numbers $\mathbb{C}$ and a complete colouring c the adjacency algebra $M_c(K)$ is semi-simple.*

For the rest of this section take $R = \mathbb{C}$. Since the only division ring over $\mathbb{C}$ is $\mathbb{C}$ itself, Wedderburn's theorem says that for the decomposition $1 = \sum_{i=1}^m \varepsilon_i$ of 1 into minimal central orthogonal idempotents and $M_i = M_c \varepsilon_i$ there is a decomposition $M_c = \bigoplus_{i=1}^m M_i$ of M_c into a direct sum of full matrix algebras M_i over $\mathbb{C}$. If M_i consists of $e_i \times e_i$ matrices, then the minimal irreducible left (or right) M_i-modules have dimension e_i and character ζ_i, say. The vector space $\mathbb{C}^V$ on which M_c acts decomposes as $\mathbb{C}^V = \bigoplus_{i=1}^m \varepsilon_i \mathbb{C}^V$ where $\varepsilon_i \mathbb{C}^V$ is a direct sum of, say, z_i copies of the irreducible

M_i-module with character ζ_i. If $\mathbb{C}^V$ has character ζ then $\zeta=\sum_{i=1}^m z_i\zeta_i$ and equating degrees gives $n=|V|=\sum_{i=1}^m z_ie_i$ and $|c|=\sum_{i=1}^m e_i^2$. Clearly $z_i\geqq 1$ for each i since the representation of M_c in $M_V(\mathbb{C})$ is faithful.

By the Noether-Skolem theorem there is an invertible matrix $U\in M_V(\mathbb{C})$ such that for all $A\in M_c$,

$$U^{-1}AU=\operatorname{diag}(D_1(A),\cdots,\underbrace{D_i(A),\cdots,D_i(A)}_{\text{multiplicity } z_i},\cdots,D_m(A))$$

is a block diagonal matrix with $D_i(A)$ affording $\zeta_i(A)$.

For generic A, $\det D_i(A)$ is irreducible as follows. Since $D_i(A)=(d_{rs})$ is generic for M_i every entry is distinct and independent of the others. Let $\det D_i(A)=fg$ be a nontrivial factorisation and $x=d_{11}$. Without loss of generality $\deg_x f=1$ and $\deg_x g=0$. Choose entry y with $\deg_y g=1$ and $\deg_y f=0$. As fg contains a term which is a multiple of xy we may assume $y=d_{22}$ by row and column interchanges. Take $d_{12}=d_{21}=1$, $d_{rs}=0$ otherwise for $r\neq s$, and $d_{rr}=1$ for $r>2$. Then $\det D_i(A)$ specialises to $xy-1$ which fails to factorise in the required way. So $\det D_i(A)$ is irreducible.

Thus, if A is generic then $\prod_{i=1}^m \det D_i(A)^{z_i}$ is the factorization of $\det A$ into its irreducible factors. Hence $\det A$ determines the e_i and $z_i(\geqq 1)$ uniquely. They in turn determine $M_c\subseteq M_V(\mathbb{C})$ up to similarity.

Conversely, to obtain a determinant for a given similarity class, pick a matrix representation containing a generic matrix whose distinct entries are linearly independent and which generates the algebra.

THEOREM 4.2. *For a complete colouring c, $M_c(\mathbb{C})$ is determined up to similarity by the determinant of a generic matrix, and conversely.*

Warning 4.3. R. Mathon [6] has some regular graphs on 25 vertices which yield complete 3-colourings that are not isomorphic but have similar adjacency algebras over $\mathbb{C}$. These also appear in [10] and seem to have been computed independently by several people.

Consider next maps $A: c\to\mathbb{C}: c_i\mapsto a_i$ from the colours $c_i(1\leqq i\leqq r)$ of c into $\mathbb{C}$. Let $\mathbb{C}^c$ denote the set of such maps. If for each $(i,j)\in V\times V$ we are given k such that $c(i,j)=c_k$ then the structure of A as an adjacency matrix is given by $a_{ij}=a_k$ and we obtain a map $\det c:\mathbb{C}^c\to\mathbb{C}: A\to\det(a_{ij})$. Clearly, from Theorem 4.2:

COROLLARY 4.4. *For a complete colouring c,* det *c determines the adjacency algebra $M_c(\mathbb{C})$ up to similarity, and conversely.*

The maps in $\mathbb{C}^c$ form an algebra (the regular representation) isomorphic to $M_c(\mathbb{C})$ under the operations induced by the map $A=(a_{ij})\mapsto A_c$ where $A_c: c(i,j)\mapsto a_{ij}$.

THEOREM 4.5. *Suppose the partition of c into diagonal and off-diagonal colours is given for a complete colouring c. Then $M_c(\mathbb{C})$ is determined up to similarity by multiplication in $\mathbb{C}^c$ defined on its natural basis.*

Proof. By Corollary 4.4 it suffices to reconstruct $\det c$. Multiplication can be described giving the intersection numbers n_{ijk} such that $A_cB_c=F_c\in\mathbb{C}^c$ satisfies $f_j=\sum_{i,k} n_{ijk}a_ib_k$. If $c_i=V_i$ is a diagonal colour then a colour c_j belongs to the block $V_i\times V_i$ if, and only if, $n_{iij}=n_{jji}=1$. So $|c_i|=\sum'_{j,k} n_{jik}$ can be found where the sum is restricted to j with $c_j\subseteq V_i\times V_i$. Let $A_c\in\mathbb{C}^c$ and compute $(A^i)_c$ for $i\in\mathbb{N}$. Then each trace $\operatorname{Tr}(A^i)$ can be calculated using $\operatorname{Tr} A=\sum'_i |c_i|a_i$ where the sum is over diagonal classes. Newton's formulae then yield $\det A$ and we obtain $\det c$.

This theorem is implicit in Higman [4, § 5]. There is a partial converse to the above which is given in [9].

Remark 4.6. Det c provides the spectrum of a graph, and, by virtue of the proof of 4.5, it follows that the intersection numbers determine equivalence classes of graphs which are at least as fine as those given by the spectrum.

5. The regular representation of the adjacency algebra. The adjacency ring $M_c = M_c(R)$ of a complete colouring c is the set of maps $V \times V \to R$ which are constant on colours of c with suitable multiplication. This gives the standard representation of M_c as a ring of matrices operating on R^V. The regular representation is given by considering M_c as the set R^c of maps from the set of colours of c to R. It is obtained as a ring of matrices as follows.

DEFINITION 5.1. For a colouring c and $A_V = (a_{ij}) \in M_V(R)$ the standard (i.e. adjacency matrix) representation of $A \in M_c$, define the matrix $A_c = (a_{lm})$ with entries indexed by colours l, m of c by

$$a_{lm} = |l|^{-1/2}|m|^{-1/2} \sum_{i,j,t \in V} |(t, i) \cap l||(t, j) \cap m| a_{ij}.$$

These matrices A_c acting on R^c give the *regular representation of* M_c.

Remark 5.2. Higman [4] makes a slightly different definition for complete colourings, namely

$$a'_{lm} = \sum_{i \in V} |(t, i) \cap l| a_{ij}$$

where $(t, j) \in m$. This is independent of the choice of $t, j \in V$ by virtue of Lemma 3.1(ii). Summing over all such (t, j) to incorporate this symmetry yields $a'_{lm} = |m|^{-1} \sum_{t,i,j \in V} |(t, i) \cap l||(t, j) \cap m| a_{ij}$. Thus

$$a_{lm} = |l|^{-1/2}|m|^{1/2} a'_{lm}.$$

In other words, rows and columns have been multiplied by certain factors.

PROPOSITION 5.3. *Let c_R be the colouring defined on a set of $|c|$ vertices by the regular representation of a colouring $c \geqq c_I$. Then c_R is symmetric* (*respectively, totally symmetric*) *if and only if c is. Also, $c_R \geqq c_I$.*

Proof. First observe that if $l \in c$ and d is the diagonal colour in the same row as l then $a_{dl} = |d|^{-1/2}|l|^{-1/2} \sum_{i:(i,i) \in d} \sum_{j:(i,j) \in l} a_{ij} = u a_{ij}$ for any $(i, j) \in l$ and some constant $u \neq 0$ dependent only on l. Hence the map $(a_{ij}) \mapsto (a_{lm})$ is one-one. It now suffices to notice from the formula that $(a_{ij})^T \mapsto (a_{lm})^T$.

Finally, a_{ll} contains a nonzero multiple of a_{ii} if $i, t \in V$ are chosen with $(t, i) \in l$, but for no i can a_{ii} appear in a_{lm} if $l \neq m$. So $c_R \geqq c_I$.

Examples 5.4. The following are generic adjacency matrices paired with their regular matrix images:

$$\begin{bmatrix} a & c & c & b & d & d \\ c & a & c & d & b & d \\ c & c & a & d & d & b \\ b & d & d & a & c & c \\ d & b & d & c & a & c \\ d & d & b & c & c & a \end{bmatrix} \text{ and } \begin{bmatrix} a & b & \sqrt{2}c & \sqrt{2}d \\ b & a & \sqrt{2}d & \sqrt{2}c \\ \sqrt{2}c & \sqrt{2}d & a+c & b+d \\ \sqrt{2}d & \sqrt{2}c & b+d & a+c \end{bmatrix};$$

$$\begin{bmatrix} a & b & d & d \\ b & a & d & d \\ g & g & e & f \\ g & g & f & e \end{bmatrix} \text{ and } \begin{bmatrix} a & b & \sqrt{2}d & 0 & 0 & 0 \\ b & a & \sqrt{2}d & 0 & 0 & 0 \\ \sqrt{2}g & \sqrt{2}g & e+f & 0 & 0 & 0 \\ 0 & 0 & 0 & e & f & \sqrt{2}g \\ 0 & 0 & 0 & f & e & \sqrt{2}g \\ 0 & 0 & 0 & \sqrt{2}d & \sqrt{2}d & a+b \end{bmatrix}.$$

LEMMA 5.5. *If c is complete, the map $(a_{ij}) \mapsto (a_{lm})$ from the standard to the regular representation is an R-module ring monomorphism.*

Proof. The property for addition is clear. Suppose (a_{ij}) and (a'_{ij}) are two adjacency matrices with images (a_{lm}) and (a'_{lm}). Using the formula in Remark 5.2, $(b_{lm}) = (a_{lm})(a'_{lm})$ has $b_{lm} = \sum_{n\in c} a_{ln}a'_{nm} = \sum_{n\in c} \{|l|^{-1/2}|n|^{1/2}\sum_{i\in V} |(t, i) \cap l|a_{ij}\}\{|m|^{-1/2}|n|^{1/2} \times \sum_{k\in V} |(t, k) \cap m|a'_{jk}\}$ where $(t, j) \in n$. Summing over all $(t, j) \in n$ and all $n \in c$ yields $b_{lm} = |l|^{-1/2}|m|^{-1/2}\sum_{i,k,t\in V} |(t, i) \cap l||(t, k) \cap m| \sum_{j\in V} a_{ij}a'_{jk}$ which is the lm-entry of the image of $(a_{ij})(a'_{ij})$. As in Proposition 5.3, the map is one-one.

THEOREM 5.6. *Let $V_1, V_2, \cdots, V_t$ be the diagonal colours of a complete colouring c. Suppose n_i is the number of colours in $V_i \times V$, so that $|c| = \sum_{i=1}^{t} n_i$. Then the matrices giving the regular representation of $M_c(\mathbb{C})$ are block diagonal with blocks of size $n_i \times n_i$ for $1 \leqq i \leqq t$.*

Proof. Suppose $l, m \in c$ with $l \subseteq V_i \times V$. If $m \subsetneq V_i \times V$ then $a_{lm} = 0$ because $|(t, i) \cap l| = 0$ whenever $(t, j) \in m$. The closure under the transpose map described in Proposition 5.3 ensures that $a_{ml} = 0$ also. This establishes the block diagonal nature of the matrices, each block being indexed by the n_i colours in $V_i \times V$ for its rows and columns.

Any map $f: V \to W$ of finite sets can be used to obtain a colouring on fV from a colouring on V. In terms of graphs the map f replaces each set $f^{-1}(w)$ of vertices in V by a single vertex $w \in fV$. In practice, f can be viewed as an equivalence relation on V which identifies various vertices.

DEFINITION 5.7. (i) For subsets S, T of V we define $c(S, T) = \{c(s, t) | s \in S, t \in T\}$, counting each $c(s, t)$ with the appropriate multiplicity.

(ii) If $f: V \to W$ is a map of finite sets and c a colouring on V then fc is the colouring on fV defined by $fc((i, j)) = c(f^{-1}i, f^{-1}j)$.

(iii) In case f is written as an equivalence relation $\sim$ on V (mapping V to $\tilde{V}$) we write $\tilde{c}$ for the colouring fc on $\tilde{V}$.

LEMMA 5.8. *If $A = (a_{ij})$ is a generic matrix for the colouring c on V and $\tilde{/}$ is an equivalence relation on V then $\tilde{c}$ has adjacency marix $\tilde{A}$ with entries*

$$\tilde{a}_{uv} = |u|^{-1/2}|v|^{-1/2} \sum_{i\in u} \sum_{j\in v} a_{ij} \quad \textit{for } u, v \in \tilde{V}.$$

Note, however, that $\tilde{A}$ need not be generic for $\tilde{c}$.

Proof. Put $a_{uv} = \sum_{i\in u} \sum_{j\in v} a_{ij}$ for $u, v \in \tilde{V}$. Then (a_{uv}) is an adjacency matrix for $\tilde{c}$. For any linear function $f = \sum_{i,j\in V} \lambda_{ij}a_{ij}$ of the a_{ij}'s let $\|f\| = \sum_{i,j\in V} |\lambda_{ij}|$. Then $\|a_{uv}\| = |u||v|$ and $\|\tilde{a}_{uv}\| = |u|^{1/2}|v|^{1/2}$. Hence $a_{uv} = a_{xy}$ if, and only if, $\tilde{a}_{uv} = \tilde{a}_{xy}$. So $(\tilde{a}_{uv})$ is also an adjacency matrix.

THEOREM 5.9. *Define an equivalence relation $\sim$ on V by $i \sim j$ if, and only if, $c(1, i) = c(1, j)$ where c is a complete colouring. Let $A \mapsto \tilde{A}$ be the map $M_c(R) \to M_{\tilde{c}}(R)$ given in Lemma 5.8. Then $\tilde{A}$ is the first block of A_c in the regular representation when the indices are paired $c(1, i)$ with $\tilde{i}$.*

Proof. Let (a_{ij}) be an adjacency matrix for c, $(\tilde{a}_{\tilde{i}\tilde{j}})$ the image under $\tilde{/}$ and (a_{lm}) the first block of the regular matrix.

Write $\tilde{i}$ instead of $c(1, i)$ to index the regular matrix block. So

$$\begin{aligned} a_{\tilde{i}\tilde{j}} &= |c(1, j)|^{1/2}|c(1, i)|^{-1/2} \sum_{i\in \tilde{i}} a_{ij} \\ &= |\tilde{j}|^{-1}|c(1, j)|^{1/2}|c(1, i)|^{-1/2} \sum_{i\in \tilde{i}, j\in \tilde{j}} a_{ij} = \tilde{a}_{\tilde{i}\tilde{j}} \end{aligned}$$

since $|V_1||\tilde{i}| = |c(1, i)|$ where V_1 is the first diagonal colour.

Remarks 5.10. Naturally, Theorem 5.9 is the fastest way to obtain the regular representation. Moreover, this representation is independent of the vertex numbering. By the definitions, it is entirely determined by the intersection numbers, and conversely.

6. Isomorphisms.

DEFINITION 6.1. Let c and d be colourings on V and W respectively. An *isomorphism* from c to d is a bijection $f\colon V \to W$ such that $fc = d$ in the notation of Definition 5.7. If, in addition, $\phi : c \to d$ is a bijection between the colours of c and d then f is called a *ϕ-isomorphism* if f induces ϕ on the colours. In particular, if $V = W$ and $c = d$ then an isomorphism is an automorphism and vice-versa; and when ϕ is the identity, then a ϕ-isomorphism is just a strict automorphism. In general, f will map the diagonal colours of c onto the diagonal colours of d and applying the transpose to colours commutes with the map f induces on colours. We will require ϕ to have these properties.

If c and d arise from two graphs then ϕ is usually the map which matches properties of one graph with those of the other. Then the existence of a ϕ-isomorphism from c to d is equivalent to the graphs being isomorphic. By viewing $f\colon V \to W$ as a re-naming of subscripts, we have (cf. Lemma 2.2(i)) the next lemma.

LEMMA 6.2. *Let $f\colon V \to W$ be injective and c, d colourings on V. Then*

(i) *a generic matrix for c is generic for fc;*

(ii) $f(cd) = f(c)f(d)$; $f(c + d) = f(c) + f(d)$; $f(c^T) = (fc)^T$;

(iii) $f(\bar{c}) = \overline{fc}$.

DEFINITION 6.3. Let c, d be symmetric colourings $\geqq c_I$. Suppose $\phi : c \to d$ is a bijection of colours which restricts to a bijection between the diagonal colours and which commutes with the transpose map. There is an induced R-module isomorphism $\Phi : R^c \to R^d$ of regular representations. If Φ commutes with multiplication, then it extends to a map $\Phi^2 : R^{c^2} \to R^{d^2} : AB \to \Phi(A)\Phi(B)$ for $A, B \in R^c$. This yields a bijection $\phi^2 : c^2 \to d^2$. Equivalently, if for all $i, j \in V$ there are $r, s \in W$ with $\{\phi c(i, t) \times \phi c(t, j) | t \in V\} = \{d(r, t) \times d(t, s) | t \in W\}$ then ϕ has a natural refinement to a bijection $\phi^2 : c^2 \to d^2$, namely $\phi^2 c^2(i, j) = d^2(r, s)$. Note, however, that ϕ^2 can be found from the multiplications $R^c \times R^c \to R^{c^2}$ and $R^d \times R^d \to R^{d^2}$ without referring back to the standard representation. In the same way, it may be possible to define $\phi^r : c^r \to d^r$ for all $r > 0$. Then iteratively one obtains a bijection $\bar{\phi} : \bar{c} \to \bar{d}$ inducing $\bar{\Phi} : R^{\bar{c}} \to R^{\bar{d}}$. If this is an R-ring isomorphism, i.e. preserves multiplication, or equivalently, $\bar{\phi}^2 = \bar{\phi}$, then we say $\bar{\phi}$ is *complete.*

There is an obvious correspondence between adjacency matrices $A = (a_{ij}) \in M_c(R)$ and $B = (b_{rs}) \in M_d(R)$ when there is a bijection $\phi : c \to d$ namely that with $a_{ij} = b_{rs}$ whenever $\phi c(i, j) = d(r, s)$. Again, let $\Phi : M_c(R) \to M_d(R)$ denote the map. We say c and d are *cospectral* (under ϕ) if, and only if, $\det A = \det \Phi A$ for all $A \in M_c(R)$, (i.e. if, and only if, $\det c = \det d_0\Phi$) and ϕ gives a bijection between diagonal colours.

THEOREM 6.4. *Suppose $f\colon V \to W$ is a ϕ-isomorphism of the colourings c, d. Then there is a natural way of refining ϕ to a complete bijection $\bar{\phi} : \bar{c} \to \bar{d}$ independently of f so that f is a $\bar{\phi}$-isomorphism from $\bar{c}$ to $\bar{d}$.*

Proof. $\bar{c}$ and $\bar{d}$ are isomorphic under f by 6.2(iii). Since $f(AB) = f(A)f(B)$ for all $A, B \in M_c(R)$, $\phi^2 : c^2 \to d^2$ may be defined by $\phi^2 c^2(i, j) = d^2(fi, fj) \equiv \{d(fi, t) \times d(t, fj) | t \in W\} = \{\phi c(i, t) \times \phi c(t, j) | t \in V\}$. So $\bar{\phi}$ is obtained by iteration, and it is complete.

THEOREM 6.5. *Suppose ϕ is a bijection between the colours of c and d, and ϕ can be refined to a complete bijection $\bar{\phi} : \bar{c} \to \bar{d}$ of colours. Then $M_{\bar{c}}(\mathbb{C})$ and $M_{\bar{d}}(\mathbb{C})$ are similar, and c and d are cospectral (under ϕ). If $V_1, \cdots, V_r$ are the diagonal colours of $\bar{c}$ and $\phi V_i = V_i$ for all i then there is a unitary matrix U, necessarily block diagonal under the partition given by the V_i's, such that $U\Phi A = AU$ for all adjacency matrices A of $\bar{c}$. Here A and ΦA have identical characteristic polynomials. There is also a block diagonal matrix $U_{\mathbb{Q}}$ with rational entries such that $U_{\mathbb{Q}}\Phi A = AU_{\mathbb{Q}}$ for all such matrices A. Moreover, U and $U_{\mathbb{Q}}$ may be chosen to have row and column sums equal to* 1.

Proof. The regular representations are identical except for the indexing by $\bar{c}$ or $\bar{d}$. Now apply Theorems 4.2 and 4.5.

There is a unitary matrix $U \in M_v(\mathbb{C})$ independent of the choice of A such that $\Phi A = U^{-1}AU$. Decomposing into blocks under the diagonal colours gives $\sum_t U_{it}\Phi A_{tj} = \sum_t A_{it}U_{tj}$. If A is a generic matrix whose elements are independent of those in U, then equating terms from the block A_{ij} yields $U_{ii}\Phi A_{ij} = A_{ij}U_{jj}$ and $U_{it} = 0$ for $t \neq i$. Hence U is block diagonal with unitary diagonal blocks.

The matrix $U_{\mathbb{Q}}$ is obtained by observing that without loss of generality U has algebraic number entries and then summing $U\Phi A = AU$ over all conjugates. $A = J = \Phi A$ gives the row and column sum property.

ALGORITHM 6.6. The graph isomorphism problem is that of finding a permutation matrix U such that $U\Phi A = AU$ for corresponding adjacency matrices A, ΦA of two graphs. This has been translated into finding a permutation $f: V \to W$ of the vertex sets which is a ϕ-isomorphism of the appropriate colourings c, d. By Theorem 6.4 there must be a complete bijection $\bar{\phi}: \bar{c} \to \bar{d}$. A basic check for isomorphism therefore involves iteratively forming c^{2^i}, d^{2^i} and ϕ^{2^i} to obtain $\bar{\phi}: \bar{c} \to \bar{d}$. This establishes that the regular representations are the same so that the standard representations by adjacency algebras are similar and the graphs co-spectral. The partitioning of the vertices via the diagonal colours serves to restrict the possible permutations if the graphs are isomorphic and standard techniques (see [7]) enable a tree of completions to be used to yield isomorphisms.

To construct the completions for two graphs and the map between their colours, represent the graphs by adjacency matrices with integer entries that are equal for edges if and only if they have identical labels in the graphs. These entries can be chosen in the range 1 to n^2 for $n = |V|$. If this can be done in $O(n^3)$ time then the $2\log_2 n$ squarings lead to an $O(n^3 \log n)$ time bound on completion, assuming that integers in the range $1 .. n^2$ can be accessed and compared in unit time. First of all, observe that even bubble sort will sort the elements of each row into order in $O(n^3)$ time, providing a permutation to reorder the elements as they appear in the row, and information about repeated elements. The same applies to columns.

Each of the n^2 elements of the square is given by a formal dot product of a row with a column. The information about how to sort both row and column must be combined to sort the n-tuple in linear time. For each distinct value in the row we have a series of adjacent spaces in the final sorted n-tuple into which terms containing that value will be placed. Assign a pointer for each such value, setting it to the first such place which is empty. Now use the column order to take each term in turn, placing it according to the corresponding row pointer, and incrementing that pointer. This sorts the n-tuple in $O(n)$ time.

The other part of the squaring procedure involves renumbering entries to obtain new numbers which are equal if and only if the corresponding sorted n-tuples are equal. This is done by renumbering using the first term, then taking the new numbering with the second term, and so on. Thus, all n-tuples must be sorted first, requiring $O(n^3)$ space to be available. Each n-tuple is represented by a vector of $2n$ integers in the range $1 .. n^2$. It suffices to show how to incorporate the first element of each into the new numbering in $O(n^2)$ time to achieve the $O(n^3)$ time requirement for squaring.

Generally, a unique numbering is obtainable for m ordered pairs of integers in the range $1 .. k$ in $O(\max(k, m))$ time. We apply this to pairs given by the current matrix numbering with the next element in each vector. The numbering is achieved by setting k list head pointers to zero and scanning each pair to set up linked lists connecting pairs with the same initial element; then each list is scanned to form sublists divided according to the second element; finally the lists are scanned again, assigning a new number of each sublist: $O(k+m)$ time.

The above process must be carried out simultaneously on both graphs to ensure common renumberings. If at any point a discrepancy arises—differing multiplicities between the two adjacency matrices—then the graphs cannot be isomorphic, and indeed, eventually there are no numbers in common in the completions. If the completions do agree then the graphs are similar if not actually isomorphic.

REFERENCES

[1] N. Biggs, *Automorphic graphs and the Krein condition*, Geometriae Dedicata, 5 (1976), pp. 117-127.

[2] R. C. Bose and T. Shimamoto, *Classification and analysis of partially balanced incomplete block designs with two associate classes*, J. Amer. Stat. Assoc., 47 (1952), pp. 151-184.

[3] G. Frobenius, *Über die Primfactoren der Gruppendeterminante*, Collected Works, Vol. 3.

[4] D. G. Higman, *Coherent configurations*, Geometriae Dedicata, 4 (1975), pp. 1-32.

[5] C. R. Johnson and M. Newman, *A note on cospectral graphs*, J. Comb. Theory B, 28 (1980), pp. 96-103.

[6] R. Mathon, *Sample graphs for graph isomorphism testing*, Technical Report, Dept. Computer Science, Univ. Toronto, to appear.

[7] R. C. Read and D. G. Corneil, *The graph isomorphism disease*, J. Graph. Theory, 1 (1977), pp. 339-363.

[8] J. J. Seidel, *Strongly regular graphs*, Surveys in Combinatorics, LMS Lecture Notes 38, B. Bollobás, ed., Cambridge Univ. Press, Cambridge, 1979, pp. 157-180.

[9] C. D. Walter, *A note on intersection numbers of coherent configurations*, J. Comb. Theory B, 31 (1983), pp. 201-204.

[10] B. Weisfeiler, ed., *On Construction and Identification of Graphs*, Lecture Notes in Mathematics 558, Springer, Berlin, 1976.

[11] G. Gati, *Further annotated bibliography on the isomorphism disease*, J. Graph Theory, 3 (1979), pp. 95-109.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

004

UPDATING *LU* FACTORIZATIONS FOR COMPUTING STATIONARY DISTRIBUTIONS*

R. E. FUNDERLIC† AND R. J. PLEMMONS‡

Abstract. The computation of stationary probability distributions for Markov chains is important in the analysis of many models in the mathematical sciences, such as queueing network models, input-output economic models and compartmental tracer analysis models. These computations often involve the solution of large-scale homogeneous linear equations by Gaussian elimination, where A is a Q-matrix, i.e., $A=(a_{ij})$ is irreducible, $a_{ij} \leqq 0$ for all $i \neq j$ and has zero column sums. The stationary distributions are the components of the unique solution vector x of positive components whose sum is one. Stable direct methods for computing x by triangular factorization $A=LU$ have received considerable attention recently and the purpose of this paper is to provide a stable method for updating the factors L and U in $O(n^2)$ flops in the case where a column of A is modified. Updating formulas are derived here using an approach similar to that for updating the Cholesky factor of a symmetric positive definite matrix after the addition of a rank one matrix. The algorithm is effective more generally for any matrix that has a stable LU factorization and for which the updated matrix has a stable LU factorization. An error analysis for thw LU update algorithm is outlined along the lines of that given for the Cholesky update by Fletcher and Powell. Details of the algorithm based on the error analysis and other considerations are given.

AMS(MOS) subject classifications. 65F05, 15A23, 15A51, 68C15, 60J20

1. Introduction.

1.1. Background.

Consider an $n \times n$ real irreducible matrix $A=(a_{ij})$ with $a_{ij} \leqq 0$ for all $i \neq j$ and with $\sum_{i=1}^{n} a_{ij}=0$, for $1 \leqq j \leqq n$. Adopting the terminology in Rose (1984), and elsewhere, we call such matrices *Q-matrices.* They arise in several areas, including the analysis of queueing networks (see, e.g., Kaufman (1983)), in the analysis of compartmental biological models (see, e.g., Funderlic and Mankin (1981)), in the input-output analysis of economic models (see, e.g., Berman and Plemmons (1981, Chap. 9), and even in the least squares adjustment of geodetic networks (see, e.g., Brandt (1983)). Q-matrices form a subclass of the widely studied class of singular irreducible M-matrices and thus they possess several important properties (see, e.g., Berman and Plemmons (1979, Chap. 6). Of particular interest to us here is that they possess LU factorizations where L and U are M-matrices. If L is chosen with ones down its main diagonal, the factorization is unique and $u_{nn}=0$ (see Funderlic and Mankin (1981)).

Assume that a Q-matrix A is given in its LU factored form

$$A=LU=\sum l_i u_i',$$

where L and U^T are lower triangular matrices with L having ones on its diagonal, the columns of L denoted by l_i and the rows of U by u_i'. Such factorizations can be computed in a stable way by Gaussian elimination since Q-matrices are diagonally dominant. In particular, the elements of A do not grow at all in magnitude during the factorization process (see, e.g., Funderlic and Plemmons (1981)). The stability of the LU factorization is not affected by a symmetric permutation P^TAP of A, where P is

* Received by the editors January 5, 1984, and in revised form August 6, 1984. This research was sponsored by the Applied Mathematics Sciences subprogram of the Office of Energy Research, U.S. Department of Energy under contract DE-AC05-840R21400 with Martin Marietta Energy Systems, Inc.

† Mathematics and Statistics Research Section, Engineering Physics and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831.

‡ Departments of Mathematics and Computer Science, North Carolina State University, Raleigh, North Carolina 27695-8205. This author's research was sponsored in part by the National Science Foundation under grant MCS-82-19500 and the Air Force Office of Scientific Research under grant AFOSR-83-0255.

generally chosen to reduce the fill-in in computing L and U where A is large and sparse. This paper is concerned with computing in an efficient and stable way the LU factorization of the modified matrix

$$\tilde{A} = A - ye_k^T, \tag{1.1}$$

where y is such that $\tilde{A}$ remains a Q-matrix. Here the vector e_k denotes a unit axis vector with one as its kth component, zeros elsewhere, and y denotes a column vector. Thus A is modified only in its kth column. The LU factorization of $\tilde{A}$ will be denoted by $\tilde{L}\tilde{U}$ with the columns of $\tilde{L}$ denoted by $\tilde{l}_i$ and the rows of $\tilde{U}$ by $\tilde{u}_i'$. Because L and U are M-matrices their sign pattern implies that if $a_{ij} \neq 0$ and $j > i$, then $u_{ij} \neq 0$ and for $j < i$ if $a_{ij} \neq 0$, then $l_{ij} \neq 0$. Therefore, if only the nonzero entries of the kth column of A are modified, then $\tilde{L}$ and $\tilde{U}$ can have nonzero elements only where L and U have nonzero elements. Thus the same data structure storage scheme for L and U can also be used to store the modified factors $\tilde{L}$ and $\tilde{U}$. Of course $\tilde{L}$ and $\tilde{U}$ can be calculated in $O(n^3)$ operations by Gaussian elimination on A, but we will be concerned with updating L and U to compute $\tilde{L}$ and $\tilde{U}$ in only $O(n^2)$ operations, in the spirit, of e.g., Bennett (1965), Fletcher and Powell (1974), Gill, Golub, Murray and Saunders (1974), or Gill and Murray (1977).

1.2. Modification of one column. Wilkinson (1977) has said that the Sherman and Morrison formula

$$(A - yv^T)^{-1} = A^{-1} - [1/(v^TA^{-1}y - 1)]A^{-1}yv^TA^{-1} \tag{1.2}$$

where y and v are column vectors is "...perhaps the most widely used result in numerical linear algebra and linear programming..." The earliest appearance of this formula is probably Duncan (1944, p. 666). The solution to the nonsingular modified system of equations

$$(A - yv^T)\tilde{x} = b, \tag{1.3}$$

when the solutions to $Ax = b$ and $Az = y$ are known, follows from (1.2) and is given by

$$\tilde{x} = x - (v^Tx/(v^Tz - 1))z. \tag{1.4}$$

Note that for nonsingular $A - yv^T$, $v^Tz \neq 1$ since otherwise z is in the null space of $A - yv^T$.

In Funderlic and Mankin (1981) it was shown for Q-matrices that the solution to a modified homogeneous system can also be obtained from (1.4). In fact if y and b are any vectors in the range of any matrix A and the solutions to $Ax = b$ and $Az = y$ are known, and if b is also in the range of $A - yv^T$, then a solution to (1.3) is also given by (1.4) when $v^Tz \neq 1$. A continuation of these results to the case of a general singular irreducible M-matrix was given in Harrod (1982, pp. 76-78). If $v^Tz = 1$ and $b = 0$, then a solution is given by $\tilde{x} = z$.

If $v^Tz = 1$ while $b \neq 0$, then there are two cases to consider. The underlying principle is that for consistent systems $Ax = b$, a solution is A^-b where A^- is any inner inverse of A, i.e. $AA^-A = A$. Inner inverses are often called $\{1\}$-inverses (see e.g. Ben-Israel and Greville (1974)).)

Case 1. If the range of A^T contains v, then since $Az = y$ and $v^Tz = 1$, it follows after some manipulation that any inner inverse of A, A^-, is an inner inverse of $\tilde{A}$. Thus $x = A^-b$ is a solution of the consistent system $\tilde{A}\tilde{x} = b$.

Case 2. If v is not in the range of A^T, then setting $B = I - A^- A$ where A^- is any inner inverse of A implies $v^T B \neq 0$. It follows that

$$A^- - \frac{1}{v^T BB^T v} BB^T vv^T A^-$$

is an inner inverse of $\tilde{A}$, and therefore

$$\tilde{x} = x - \frac{1}{v^T BB^T v}(BB^T vv^T)x$$

is a solution of (1.3).

Though the above discussion on updating is already more general than necessary for our purposes, more could be said. When a rank one matrix is subtracted from a given matrix, the rank of the resulting matrix may only differ by at most one from that of the given matrix. For example the condition $v^T z = 1$ implies that the rank of $\tilde{A} = A - yv^T$ is one less than the rank of A (cf. Case 1 above) if and only if v is in the range of A^T, see e.g. Householder (1964, p. 33, ex. 34). Therefore when the rank of $\tilde{A}$ is less than that of A, the null space of $\tilde{A}$ is generated by the null space of A and the vector z. This raises the question, which we shall not pursue, of characterizing completely the solution spaces for $\tilde{A}\tilde{x} = \tilde{b}$ depending on whether rank($\tilde{A}$) is one less, one more, or the same as that of A. For a further discussion of how the rank of a matrix A differs from the rank of a difference $A - S$, see Cline and Funderlic (1979).

Finally, we mention that Meyer and Shoaf (1980) have studied the general problem of updating Markov chains by updating the group generalized inverse of A. Our approach is different in that we update a triangular factorization of A instead.

1.3. Homogeneous systems. The main application we have in mind here is the solution of homogeneous systems of the form

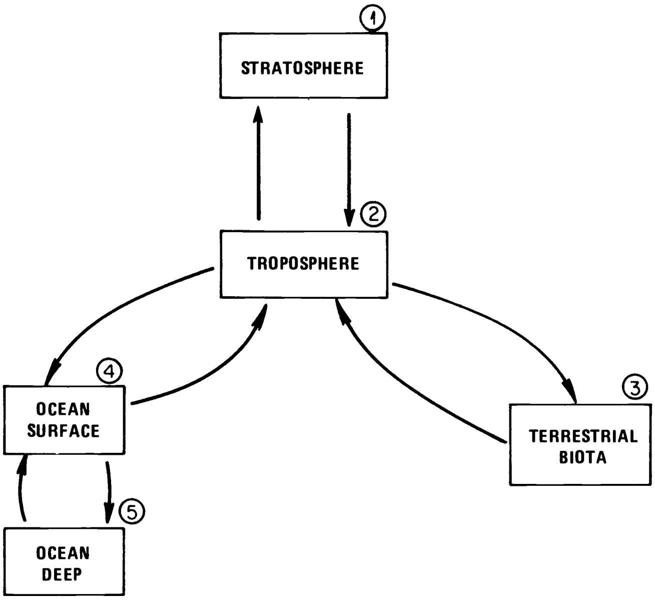
$$Ax = 0 \tag{1.5}$$

where A is a Q-matrix. Our purpose is to compute the unique *stationary probability distribution vector* $x = (x_i)$, $x_i > 0 \sum x_i = 1$, which solves the homogeneous system (1.5). Here A might be considered to be the transfer rate matrix for a finite, homogeneous, ergodic Markov process. Both iterative and direct methods for computing x have been extensively studied in the literature (see, e.g., Kaufman (1983), Funderlic and Mankin (1980), and Harrod and Plemmons (1984)). A combined direct-iterative method was studied by Funderlic and Plemmons (1984). Our interest here is in direct methods for computing x, based upon an LU factorization of A. In particular, if A is updated to $\tilde{A}$ given in (1.1), then the updated $\tilde{L}$ and $\tilde{U}$ can be used to compute the updated stationary distribution vector $\tilde{x}$ which solves

$$\tilde{A}\tilde{x} = 0.$$

A typical application in compartmental analysis would be where one needs to change the rates at which a material leaves a compartment. For example, the carbon model depicted by Fig. 1 (Gardner, Mankin and Emanuel (1980) and Funderlic and Mankin (1981)) could have a transfer rate matrix given by

$$A = \begin{bmatrix} .50 & -.090 & 0 & 0 & 0 \\ -.50 & .47 & -0.061 & -0.080 & 0 \\ 0 & -.20 & 0.061 & 0 & 0 \\ 0 & -.18 & 0 & .12 & -.0011 \\ 0 & 0 & 0 & -.04 & .0011 \end{bmatrix}.$$



GLOBAL CARBON CYCLE: MODEL 3

FIG. 1. *A carbon model.*

A simple way, suggested by Funderlic and Mankin (1981), to solve for the steady-state vector x from the system $Ax=0$ is to obtain an LU factorization of A:

$$L=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & -.53 & 1 & 0 & 0 \\ 0 & -.47 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix},$$

$$U=\begin{bmatrix} .50 & -.09 & 0 & 0 & 0 \\ 0 & .38 & -.061 & -.08 & 0 \\ 0 & 0 & .029 & -.042 & 0 \\ 0 & 0 & 0 & .04 & -.0011 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and solve the equivalent system $Ux=0$, to give

$$x^T=(.0020, .0113, .0370, .0254, .9242),$$

normalized so that $\sum x_i=1$. Let the rates out of the second compartment be changed so that the second column of $\tilde{A}$ is

$$\tilde{a}=(-.15, .49, -.18, -.16, 0)^T.$$

Then if the second column of A is denoted as a, we have

$$\tilde{A}=A-ye_2^T,$$

where

$$y=a-\tilde{a}=(.06, -.02, -.02, -.02, 0).$$

Then $Az = y$ can be efficiently solved for the vector z using the LU factorization of A: $Lp = y$, $Uz = p$. An unnormalized $\tilde{x}$ may be calculated from (1.4) from which the normalized

$$\tilde{x} = (.0038, .0127, .0374, .0253, .9209)^T$$

is calculated. Fr̈om premultiplying $Uz = p$ by e_n^T, it follows that $p_n = 0$. Thus the back-substitution for z may be started with an arbitrary z_n since $u_{nn} = 0$. If the new second column of A is modified further, new vectors $\tilde{x}$ can continue to be efficiently calculated from (1.4), with $v = e_2$.

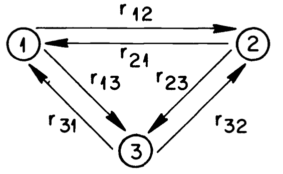
Using the terminology of C. B. Moler, we define a *flop* to be a floating point calculation consisting of one addition, one multiplication, and a little indexing, such as $c_{ij} = s + a_{ik}b_{kj}$. Then assuming that the LU factorization of A has already been carried out, $O(n^2)$ flops are required for the solution of $Az = u$ and a single flop is required for (1.4), since $v = e_k$ where the kth column of A is being changed.

1.4. Why update *L* and *U*? As long as only one column of A is to be modified in sequence, the strategy of the previous section is quite appropriate. However, if we now wish to follow a modification of column k by a modification of column $j \neq k$, then there is a problem because we now have no LU factorization of $A - ye_k^T$. Though in compartmental analysis the modification of only one column occurs frequently, in many Markov processes several different columns need to be changed. Consider the following simple queueing network given by Fig. 2, where r_{ij} is the probability that a customer exiting station i will proceed next to station j. At any instant the network is in one of the six states of Fig. 2.

For example in state 3 there is one customer at station 1 and one customer at station 3. The rate of transition from state 3 to state 5 is given by $\mu_1 r_{12}$, where μ_i is the service rate of the server at service facility i.

The network of Fig. 2 leads to the following infinitesimal generator or transition *rate* matrix (of order equal to the number of states)

$$S = \begin{bmatrix} s_{11} & \mu_1 r_{12} & \mu_1 r_{13} & 0 & 0 & 0 \\ \mu_2 r_{21} & s_{22} & \mu_2 r_{23} & \mu_1 r_{12} & \mu_1 r_{13} & 0 \\ \mu_3 r_{31} & \mu_3 r_{32} & s_{33} & 0 & \mu_1 r_{12} & \mu_1 r_{13} \\ 0 & \mu_2 r_{21} & 0 & s_{44} & \mu_2 r_{23} & 0 \\ 0 & \mu_3 r_{31} & \mu_2 r_{21} & \mu_3 r_{32} & s_{55} & \mu_2 r_{23} \\ 0 & 0 & \mu_3 r_{31} & 0 & \mu_3 r_{32} & s_{66} \end{bmatrix},$$



3 SERVICE FACILITIES (STATIONS)
EACH CONTAINING A SINGLE
EXPONENTIAL SERVER.

2 CUSTOMERS

STATE NUMBER	1	2	3	4	5	6
STATE	(2,0,0)	(1,1,0)	(1,0,1)	(0,2,0)	(0,1,1)	(0,0,2)

FIG. 2. *A simple queueing network.*

where S has nonnegative off-diagonal elements and zero *row* sums. The off-diagonal elements are determined from Fig. 2, i.e. s_{ij} is the rate of transition from state i to state j. The required stationary probability vector x is obtained by solving $x^TS=0$ or $S^Tx=0$. To determine an associated stochastic probability matrix Q, consider

$$S^T\,\Delta tx+x=x.$$

If the scalar Δt is chosen such that $\Delta t\leqq d^{-1}$, where $d=\max S_{ij}$, then

$$Q=\Delta tS+I.$$

Let $A=S^T$. Observe that if r_{ij} is changed, then 3 columns of A are altered. This again illustrates the need to update LU factorizations to obtain updated solutions of $\tilde{A}\tilde{x}=0$.

2. Updating *LU* factorizations of *Q*-matrices.

2.1. The algorithm. Bennett (1965) gave a general algorithm to update an LDU factorization of a nonsingular matrix by a rank m matrix XCY^T, where X and Y are $n\times m$ matrices. Bennett credits J. C. Butcher with pointing out to him the triangular factors of a matrix of the form $I+xx^T$. Gill et al. (1974) and Fletcher and Powell (1974) in the same volume of "Mathematics of Computation" published papers each of which considered updating an LDL^T factorization of a positive definite matrix after a rank one update σxx^T. It was pointed out that even when a Bennett update of a general matrix allowed an $\tilde{L}\tilde{D}\tilde{U}$ factorization, the factorization could be unstable (e.g. Gill et al. (1974)). In particular, the elements of $\tilde{L}$ and $\tilde{U}$ can become relatively large in magnitude. Furthermore, even for symmetric positive definite matrices the given matrix dictates branches of the algorithm to insure stability (see e.g. Fletcher and Powell (1974, p. 1074)). For Q-matrices a similar situation occurs. This will be elaborated on further in § 3.

The Gill et al. (1974) discussion is particularly compact as they point out that

$$A+\sigma zz^T=LDL^T+\sigma zz^T=L(D+\sigma vv^T)L^T=L(\bar{L}\tilde{D}\bar{L}^T)L^T,$$

where v is obtained from $Lv=z$. Alternatively Fletcher and Powell (1974) write

$$A=\sum_{i=1}^n l_id_il_i^T\quad\text{and}\quad\tilde{A}=\sum_{i=1}^n\tilde{l}_i\tilde{d}_i\tilde{l}_i^T.$$

They note that vectors $\tilde{l}_1$, z_2 and the scalars $\tilde{d}_1$ and σ_2 can be determined so that the first components of $\tilde{l}_1$ and z_2 are respectively one and zero and

$$d_1l_1l_1^T+\sigma zz^T=\tilde{d}_1\tilde{l}_1\tilde{l}_1^T+\sigma_2z_2z_2^T. \tag{2.1}$$

When equation (2.1) is subtracted from

$$\sum_{i=1}^n d_il_il_i^T+\sigma zz^T=\sum_{i=1}^n\tilde{d}_i\tilde{l}_i\tilde{l}_i^T,$$

one is left with the update problem of dimension one less:

$$\sum_{i=2}^n d_il_il_i^T+\sigma_2z_2z_2^T=\sum_{i=2}^n\tilde{d}_i\tilde{l}_i\tilde{l}_i^T.$$

Carrying out the process used to determine $\tilde{l}_1$, z_2, $\tilde{d}_1$ and σ_2, n times gives the complete factorization of $\tilde{A}$. A key observation here is that the reduced problem at each step is that of updating the triangular factorization of a symmetric positive definite matrix.

We chose the Fletcher and Powell approach to derive the formulas for our situation. To simplify the notation we write the analogue of (2.1) without subscripts and use primes for row vectors:

$$lu' + xy' = \tilde{l}\tilde{u}' + zw'. \tag{2.2}$$

Thus an LU factorization $\sum_{i=1}^{n-1} l_i u_i'$ of A is to be updated by xy'. Note that we need only sum to $n-1$ here since $u_n' = (0, \cdots, 0)$. In this section and what follows, the vector x is associated with the update rather than with a solution to $Ax = 0$. We require column vectors $\tilde{l}$ and z and row vectors $\tilde{u}'$ and w' such that $e_1^T z = w'e_1 = 0$ and $e_1^T \tilde{l} = 1$. This gives a problem of dimension one less to be updated:

$$\sum_{i=2}^{n-1} l_i u_i' + zw' = \sum_{i=2}^{n-1} \tilde{l}_i \tilde{u}_1',$$

if $l = l_1$, $u = u_1'$, $\tilde{l} = \tilde{l}_1$ and $\tilde{u}' = \tilde{u}_1'$. Further, denote the first components of x, y' and u' by ξ, η and μ. Then relation (2.2) implies

$$\tilde{u}' = u' + \xi y', \tag{2.3}$$

and

$$\tilde{\mu}\tilde{l} = \mu l + \eta x, \tag{2.4}$$

where $\tilde{\mu} = \mu + \xi\eta$. These relations follow by premultiplying relation (2.2) by e_1^T and post multiplying by e_1. If there are vectors z and w' that satisfy (2.2), then z must be a linear combination of x, and l and w' of y' and u'. Furthermore, z and w' are unique vectors up to scalar multiples. The choice

$$z = x - \xi l \tag{2.5}$$

is effective and implies from (2.2) that

$$\tilde{\mu} w' = \mu y' - \eta u'. \tag{2.6}$$

The number of multiplications can be reduced in the algorithm by observing that

$$\tilde{l} = l + \beta z \tag{2.7}$$

and

$$w' = y' - \beta \tilde{u}' \tag{2.8}$$

where $\beta = \eta/\tilde{\mu}$. Except for updating the final one-dimensional problem, $\tilde{\mu}$ must be nonzero since it is a diagonal element of $\tilde{U}$. Thus if $\beta_i = \eta_i/\tilde{\mu}_i$, an algorithm can be given from (2.3), (2.6), (2.5) and (2.4). Alternatively (2.8) can be used for (2.6), and (2.7) for (2.4). Define the vectors $x_1 = x$, $y_1' = y'$ and the scalars $\xi_i = e_i^T x_i$, $\eta_i = y_i' e_i$, $\mu_i = u_i' e_i$ and $\tilde{\mu}_i = \mu_i + \xi_i \eta_i$. The algorithm therefore takes the form for $i = 1, 2, \cdots, n-1$

$$\tilde{u}_i' = u_i' + \xi_i y_i', \tag{2.9a}$$

$$\gamma_i = \mu_i/\tilde{\mu}_i \quad \text{or} \quad \beta_i = \eta_i/\tilde{\mu}_i, \tag{2.9b, c}$$

$$y_{i+1}' = \gamma_i y_i' - \beta_i u_i' \quad \text{or} \quad y_{i+1}' = y_i' - \beta_i \tilde{u}_i', \tag{2.9d,e}$$

$$x_{i+1} = x_i - \xi_i l_i, \tag{2.9f}$$

$$\tilde{l}_i = \gamma_i l_i + \beta_i x_i \quad \text{or} \quad \tilde{l}_i = l_i + \beta_i x_{i+1}. \tag{2.9g, h}$$

We observe here that the reduced problem at each step is that of updating the LU factorization of a Q-matrix, so that a stable factorization exists at each step. This follows since after each step of Gaussian elimination on A, the unreduced part remains a Q-matrix (see, e.g., Funderlic and Mankin (1981)).

2.2. Interpretation as column exchanges and partitioning. In Gill and Murray (1977) for example, an alternate procedure is discussed for the case where A is updated by deleting and adding a column. This is known as a *column exchange.* Here, however, it is necessary to add the column at the end and the algorithm requires increasingly more steps as more columns are exchanged. In addition, some pivoting is gnerally necessary to preserve numerical stability.

The formulas (2.9) can be thought of as being associated with matrices L, $\tilde{L}$, X, U, $\tilde{U}$, and Y with the first three lower triangular and the last three upper triangular. When the kth column of A is modified and $k > 1$, then certain elements of L and $\tilde{L}$ are identical as are certain elements of U and $\tilde{U}$. To see this suppose we are given the factorization $A = LU$ and wish to update A to $\tilde{A} = A - ye_k^T$, where, as before, we assume that the updated $\tilde{A}$ remains a Q-matrix. Further, let $a_k = (s, \alpha, t)^T$ denote the kth column of A where s has dimension $k-1$, α is a scalar and t has dimension $n-k$. Also let $\tilde{a}_k = a_k - y = (u, \beta, v)^T$ be partitioned conformally with a_k. Then partitioning A, $\tilde{A}$, L, U, $\tilde{L}$ and $\tilde{U}$ conformally, we have the following block factorizations:

$$A = \begin{bmatrix} A_{11} & s & A_{12} \\ z^T & \alpha & w^T \\ A_{21} & t & A_{22} \end{bmatrix} = LU = \begin{bmatrix} L_{11} & 0 & 0 \\ l_1^T & 1 & 0 \\ L_{21} & l_2 & L_{22} \end{bmatrix} \begin{bmatrix} U_{11} & u_1 & U_{12} \\ 0 & u_{kk} & u_2^T \\ 0 & 0 & U_{22} \end{bmatrix}$$

and

$$\tilde{A} = \begin{bmatrix} A_{11} & u & A_{12} \\ z^T & \beta & w^T \\ A_{21} & v & A_{22} \end{bmatrix} = \tilde{L}\tilde{U} = \begin{bmatrix} \tilde{L}_{11} & 0 & 0 \\ \tilde{l}_1^T & 1 & 0 \\ \tilde{L}_{21} & \tilde{l}_2 & \tilde{L}_{22} \end{bmatrix} \begin{bmatrix} \tilde{U}_{11} & \tilde{u}_1 & \tilde{U}_{12} \\ 0 & \tilde{u}_{kk} & \tilde{u}_2^T \\ 0 & 0 & \tilde{U}_{22} \end{bmatrix}.$$

Now observe that

1) $\tilde{L}_{11} = L_{11}$ and $\tilde{U}_{11} = U_{11}$ since $A_{11} = L_{11}U_{11}$ uniquely,

2) $\tilde{U}_{12} = L_{11}^{-1}A_{12} = U_{12}$,

3) $\tilde{L}_{21} = A_{21}U_{11}^{-1} = L_{21}$,

4) $\tilde{l}_1^T = z^T U_{11}^{-11} = l_1^T$,

5) $\tilde{u}_2^T = w^T - l_1^T U_{12} = u_2^T$.

Then $\tilde{L}$ and $\tilde{U}$ have the block forms

$$\tilde{L} = \begin{bmatrix} L_{11} & 0 & 0 \\ l_1^T & 1 & 0 \\ L_{21} & \tilde{l}_2 & \tilde{L}_{22} \end{bmatrix}, \qquad \tilde{U} = \begin{bmatrix} U_{11} & \tilde{u}_1 & U_{12} \\ 0 & \tilde{u}_{kk} & u_2^T \\ 0 & 0 & 0 \end{bmatrix}.$$

This means that only the $(n-k)$-dimensional vector $\tilde{l}_2$, the $(k-1)$-dimensional vector $\tilde{u}_1$, the scalar $\tilde{u}_{kk}$ and the $(n-k)$-dimensional matrices $\tilde{L}_{22}$ and $\tilde{U}_{22}$ need to be recalculated in the updating process.

Bunch and Rose (1974) have also considered updating problems in conjunction with partitioning, tearing and modification schemes for general linear sparse systems.

2.3. Simple Fortran implementation. Here a simple Fortran implementation of the formulas (2.9) using (2.9a, c, e, f, h) is given by Fig. 3. Let UN and LN be Fortran arrays that denote $\tilde{U}$ and $\tilde{L}$ respectively. Further assume that UN and LN are the same as U and L initially. It is assumed that xe_k^T is added to LU so that initially $y' = e_k^T$ is in the Fortran Y vector and x in the Fortran X vector. As the algorithm progresses the appropriate leading components of X and Y are implicitly assumed zero.

```
      SUBROUTINE UPDATE(N,K,L,LN,U,UN,X,Y)
      REAL L(N,N),LN(N,N),U(N,N),UN(N,N),X(N),Y(N)
      NM1=N-1
      DO 25 I=1,NM1
        XI=X(I)
        IF(I.LT.K)GO TO 5
        UN(I,I)=U(I,I)+XI*Y(I)
        BETA=Y(I)/UN(I,I)
 5      CONTINUE
        IF(I.LT.K)UN(I,K)=U(I,K)+XI
        IP1=I+1
          DO 10 J=IP1,N
            X(J)=X(J)-XI*L(J,I)
10        CONTINUE
        IF(I.LT.K)GO TO 20
          DO 15 J=IP1,N
            UN(I,J)=U(I,J)+XI*Y(J)
            Y(J)=Y(J)-BETA*UN(I,J)
            LN(J,I)=L(J,I)+BETA*X(J)
15        CONTINUE
20      CONTINUE
25    CONTINUE
      RETURN
      END
```

FIG. 3. *A simplified update subroutine.*

Though the subroutine of Fig. 3 is not what would be implemented in a high quality subroutine, it helps illustrate several points: it is not necessary to have matrices to represent X and Y. If a flop is defined as in § 1.3, then the algorithm, ignoring lower order terms, takes between $2n^2$ and $n^2/2$ flops as k varies from 1 to n. In the final suggested algorithm the only array storage required is the original matrix plus the x and y vectors. See § 4.

3. Error analysis. Fletcher and Powell (1974) have given a complete a posteriori error analysis for the symmetric positive definite update problem of an LDL^T factorization. This analysis is quite tedious, but it does carry over for LU factorizations of diagonally dominant irreducible M-matrices.

A crucial point in the Fletcher–Powell error analysis (1974, inequality 5.16, p. 1080) is that the error term (at the ith stage) contains elements from the Cholesky factorizations of A and $\tilde{A}$. These elements are not unacceptably large since the Cholesky factorizations of A and $\tilde{A}$ are stable. Likewise the error term for the update of a Q-matrix contains elements of the stable LU factorizations of A and $\tilde{A}$. What can be inferred is that if any matrices A and $\tilde{A}$ have given stable LU factorizations and differ by a rank one matrix, then there is a stable update algorithm which will produce a stable LU factorization for $\tilde{A}$ from that of A. Following the analysis of Fletcher and Powell, when the economical (2.9e) and (2.9h) are used, $\tilde{\bar{\mu}}_i/\mu_i$ occurs in the error term where the bar indicates a calculated value. Thus when the ith diagonal element of $\tilde{U}$ is large compared with the ith diagonal element of U, unacceptable growth may occur. To offset this we use the more expensive formulas (2.9d) and (2.9g) when $\gamma = \mu/\tilde{\mu} < \frac{1}{4}$. This is in line with the Fletcher and Powell choice. When the more expensive formulas are used, $\mu/\tilde{\bar{\mu}}_i$ occurs in the error term. For most problems $\mu_i/\tilde{\mu}_i$ is seldom less than $\frac{1}{4}$, so that realistically the number of flops does not appreciably increase over that given in the simplified subroutine of Fig. 3.

When $\tilde{\mu}$ is relatively large with respect to μ, $\tilde{l}\tilde{u}'$ tends to lose the contribution of x when (2.9h) is used, and conversely when $\tilde{\mu}$ is relatively small, cancellation can occur when (2.9g) is used. Similar comments can be made with respect to xy'.

The Cholesky process can break down for a symmetric positive definite matrix A. That is, if A is poorly conditioned, a zero or negative element may appear on the diagonal of L, Wilkinson (1968). Similarly the update process may break down and much research has been done to alleviate that problem, e.g. Fletcher and Powell (1974), Gill and Murray (1977) and Dax (1983). In the symmetric positive definite case no difficulty can occur when the update is of the form σxx^T with $\sigma > 0$. The difficulty occurs when $\sigma < 0$ and has lately been called downdating. Stewart (1979) has shown that when downdating breaks down or nearly breaks down in the Cholesky process, $\tilde{L}$ is an ill-conditioned function of L and the update. In any case the methods for preventing or correcting a breakdown for Q-matrices are more complicated than that for symmetric positive definite matrices. For Q-matrices the distinction of updating and downdating cannot be made, but rather the analogue of σ can change signs at each step. Though analogous strategies to those for the symmetric case can be made for Q-matrices we will not pursue such strategies.

4. Implementation based on error analysis and storage. The purpose of this section is to indicate a way to implement the update algorithm of (2.9) with error analysis considerations in essentially n^2+2n storage locations. Again the problem is to update an $n \times n$ matrix LU by xe_k^T so that

$$\tilde{L}\tilde{U} = LU + xe_k^T.$$

The algorithm depicted by Fig. 4 assumes that the matrices L and U are stored in the

```
For     $i = 1, 2, \ldots, n-1$                                   (4.1)
            $\xi \leftarrow x_i$
            If $i < k$, then
(2.9a)          $a_{ik} \leftarrow a_{ik} + \xi$                          (4.2)
                For $j = i+1, i+2, \ldots, n$
(2.9f)              $x_j \leftarrow x_j - \xi a_{ji}$
            otherwise
                $t \leftarrow a_{ii}$
(2.9a)          $a_{ii} \leftarrow a_{ii} + \xi y_i'$
                If $a_{ii} \le 0$, then error exit                 (4.3)
                otherwise
(2.9b)              $\gamma \leftarrow t / a_{ii}$
(2.9c)              $\beta \leftarrow y_i' / a_{ii}$
                    If $\gamma < 1/4$, then
                        For $j = i+1, i+2, \ldots, n$
                            $t \leftarrow y_j'$
(2.9d)                      $y_j' \leftarrow \gamma t - \beta a_{ij}$
(2.9a)                      $a_{ij} \leftarrow a_{ij} + \xi t$
                            $t \leftarrow a_{ji}$
(2.9g)                      $a_{ji} \leftarrow \gamma t + \beta x_j$
(2.9f)                      $x_j \leftarrow x_j - \xi t$
                    otherwise
                        For $j = i+1, i+2, \ldots, n$              (4.4)
(2.9a)                      $a_{ij} \leftarrow a_{ij} + \xi y_j'$
(2.9e)                      $y_j' \leftarrow y_j' - \beta a_{ij}$
(2.9f)                      $x_j \leftarrow x_j - \xi a_{ji}$
(2.9h)                      $a_{ji} \leftarrow a_{ji} + \beta x_j$
```

FIG. 4. *The update algorithm.*

array A with the diagonal elements of U as diagonal elements of A and those of L implicitly assumed to be 1. The matrices $\tilde{L}$ and $\tilde{U}$ will overwrite L and U in the array A. In practice matrix sparsity or architecture considerations may suggest other data structures. The vectorized formulas (2.9) derived in § 2 are referenced down the left side of Fig. 4. In the formulas (2.9), x_i and y_i denoted column and row vectors whereas in Fig. 4 we start with $y' = e_k^T$ and x is the kth column of $\tilde{A} = \tilde{L}\tilde{U}$. The x vector is overwritten at each step of the algorithm and the y' vector is overwritten from the kth step on. Thus in Fig. 4 x_j and y'_j are the jth components of the current x and y' vectors. The algorithm consists of one outer loop (4.1) with three inner loops. The first inner loop is all that is executed until $i \geqq k$. The choice between the other two inner lops is made depending on how $u_{ii}/\tilde{u}_{ii}$ compares with 1/4. The outer loop only goes to $n-1$ because $u_{nn} = \tilde{u}_{nn} = 0$ for Q-matrices. If the algorithm were to be modified for nonsingular matrices, the outer loop would go to n rather $n-1$. While $i < k$, the vector $y' = e_k^T$ is unchanged so that from (2.9a), only the kth component of each of the rows of U are changed by the ith component of x. This is depicted in (4.2). Since the diagonal elements of Q-matrices are mathematically positive, an error exit is indicated at (4.3) if $\tilde{u}_{ii} \leqq 0$. The loop at (4.4) depicts the same formulas that were used in the simplified Fortran subroutine of Fig. 3.

5. A worked example. If the kth column of A, a_k, is to be changed to $\hat{a}_k$, then

$$\tilde{A} = A + (\hat{a}_k + a_k)e_k^T,$$

and therefore in the notation of § 2.1, the vector $x = \hat{a}_k - a_k$. Alternatively we can start with the kth column of A zeroed out so that $x = \hat{a}_k$. The LU factorization to be updated in the case of the zero kth column is the same as when $a_k \neq 0$ except the kth column of the upper triangular matrix U has as its kth column a zero column. Let

$$A = \begin{bmatrix} 5 & -2 & 0 & -2 & -1 \\ -2 & 5.8 & 0 & -3.2 & -4.6 \\ -1 & -2.6 & 0 & -1.2 & -2.8 \\ -1 & -.1 & 0 & 9.6 & -1.1 \\ -1 & -1.1 & 0 & -3.2 & 9.5 \end{bmatrix}.$$

The triangular factors of A are

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -.4 & 1 & 0 & 0 & 0 \\ -.2 & -.6 & 1 & 0 & 0 \\ -.2 & -.1 & -.7 & 1 & 0 \\ -.2 & -.3 & -.3 & -1 & 1 \end{bmatrix}$$

and

$$U = \begin{bmatrix} 5 & -2 & 0 & -2 & -1 \\ 0 & 5 & 0 & -4 & -5 \\ 0 & 0 & 0 & -4 & -6 \\ 0 & 0 & 0 & 6 & -6 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Let $x^T = (-2, -1, 8, -1, -4)$ and $\tilde{A} = A + xe_3^T$. Then to 3 significant digits

$$\tilde{L} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -.4 & 1 & 0 & 0 & 0 \\ -.2 & -.6 & 1 & 0 & 0 \\ -.2 & -.1 & -.242 & 1 & 0 \\ -.2 & -.3 & -.758 & -1 & 1 \end{bmatrix}$$

and

$$\tilde{U} = \begin{bmatrix} 5 & -2 & -2 & -2 & -1 \\ 0 & 5 & -1.8 & -4 & -5 \\ 0 & 0 & 6.52 & -4 & -6 \\ 0 & 0 & 0 & 7.83 & -3.25 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$X = \begin{bmatrix} -2 & 0 & 0 & 0 & 0 \\ -1 & -1.8 & 0 & 0 & 0 \\ 8 & 7.6 & 6.52 & 0 & 0 \\ -1 & -1.4 & -1.58 & 2.98 & 0 \\ -4 & -4.4 & -4.94 & -2.98 & 0 \end{bmatrix}$$

and

$$Y = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & .613 & .920 \\ 0 & 0 & 0 & 0 & 1.18 \end{bmatrix}$$

Notice that when $\tilde{A} = A + xe_3^T$, the first column of X is x and the first row of Y is e_3^T. Since the first $n-1$ column sums of L are zero, (2.9f) implies that the column sums of X are zero. For the first $k-1$ steps of the algorithm, β and γ are not calculated. However, $\beta_3 = .153$, $\gamma_3 = 0$, $\beta_4 = .0783$, and $\gamma_4 = .766$.

Initially the L and U matrices are stored in the A array as

$$\begin{bmatrix} 5 & -2 & 0 & -2 & -1 \\ -.4 & 5 & 0 & -4 & -5 \\ -.2 & -.6 & 0 & -4 & -6 \\ -.2 & -.1 & -.7 & 6 & -6 \\ -.2 & -.3 & -.3 & -1 & 0 \end{bmatrix},$$

and on return from the algorithm described in the last section the A array is overwritten by

$$\begin{bmatrix} 5 & -2 & -2 & -2 & -1 \\ -.4 & 5 & -1.8 & -4 & -1 \\ -.2 & -.6 & 6.52 & -4 & -6 \\ -.2 & -.1 & -.242 & 7.83 & -3.25 \\ -.2 & -.3 & -.758 & -1 & 0 \end{bmatrix}.$$

Acknowledgments. R. E. Funderlic wishes to thank North Carolina State University for its support and hospitality while much of this work was done. We wish to thank Professor Gene Golub for encouraging this work and Professor W. J. Stewart for the queueing example in Fig. 2.

REFERENCES

J. BENNETT [1965], *Triangular factors of modified matrices*, Numer. Math., 7, pp. 217-221.

A. BEN-ISRAEL AND T. GREVILLE [1974], *Generalized Inverses: Theory and Applications*, Wiley, New York.

A. BERMAN AND R. PLEMMONS [1979], *Nonnegative Matrices in the Mathematical Sciences*, Academic Press Series on Computer Science and Applied Mathematics, Academic Press, New York.

J. BUNCH AND D. ROSE [1974], *Partitioning, tearing and modification of linear systems*, J. Math. Anal. Appl., 48, pp. 574-583.

A. BRANDT [1983], *Algebraic multigrid theory*, preprint.

R. CLINE AND R. FUNDERLIC [1979], *The rank of a difference of matrices and associated generalized inverses*, Linear Alg. and Appl., 24, pp. 185-215.

A. DAX [1983], *A diagonal modification for the downdating algorithm*, SIAM J. Sci. Stat. Comp., 4, pp. 85-93.

W. DUNCAN [1944], *Some devices for the solution of large sets of simultaneous equations*, Philos. Mag., 7, pp. 660-670.

R. FLETCHER AND M. POWELL [1974], *On the modification of LDL^T factorizations*, Math. Comp., 28, pp. 1067-1087.

R. FUNDERLIC AND J. MANKIN [1981], *Solution of homogeneous systems of equations arising from compartmental models*, SIAM J. Sci. Stat. Comp., 2, pp. 375-383.

R. FUNDERLIC AND R. PLEMMONS [1981], *LU decompositions of M-matrices by elimination without pivoting*, Linear Alg. and Appl., Vol. 41, 99-110.

R. FUNDERLIC AND R. PLEMMONS [1984], *A combined direct-iterative method for certain M-Matrix linear systems*, this Journal, 5, pp. 33-42.

R. GARDNER, J. MANKIN AND W. EMANUEL [1980], *A comparison of three carbon models*, Ecol. Modelling, 8, pp. 313-332.

P. GILL, G. GOLUB, W. MURRAY AND M. SAUNDERS [1974], *Methods for modifying matrix factorizations*, Math. Comp., 28, pp. 505-535.

P. GILL AND W. MURRAY [1977], *Modification of matrix factorizations after a rank one change*, Proc. Conference on The State of the Art in Numerical Analysis at the University of York, Academic Press, New York, pp. 55-83.

W. HARROD [1982], *Rank modification method for certain singular systems of linear equations*, Ph.D. Dissertation, Univ. Tennessee, Knoxville.

W. HARROD AND R. PLEMMONS [1984], *Comparison of some direct methods for computing stationary distributions of Markov chains*, SIAM J. Sci. Stat. Comp., 5, pp. 453-469.

A. S. HOUSEHOLDER [1964], [1975], *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York; Dover, New York.

L. KAUFMAN [1983], *Matrix models for queueing problems*, SIAM J. Sci. Stat. Comp., 4, pp. 525-552.

C. MEYER AND J. SHOAF [1980], *Updating finite Markov chains by using techniques of group matrix inversion*, J. Statist. Comput. Simul., 11, pp. 163-181.

D. ROSE [1984], *Convergent regular splittings for singular M-matrices*, this Journal, 5, pp. 133-144.

G. STEWART [1979], *The effects of rounding error on an algorithm for downdating a Cholesky factorization*, J. Inst. Math. Appl., 23, pp. 203-213.

J. WILKINSON [1968], *A priori error analysis of algebraic processes*, Proc. International Congress Mathematics, Mir, Moscow, pp. 629-639.

——— [1977], *Some recent advances in numerical linear algebra*, Proc. Conference on The State of the Art in Numerical Analysis at the University of York, Academic Press, New York, p. 4.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

THE GEOMETRY OF m-SEQUENCES: THREE-VALUED CROSSCORRELATIONS AND QUADRICS IN FINITE PROJECTIVE GEOMETRY*

RICHARD A. GAMES†

Abstract. Hyperplanes H and sets H^r of $PG(n-1, 2)$ are identified with pairs of binary m-sequences of span n. If H^r is a quadric, then a three-valued periodic crosscorrelation function between the m-sequences results. Conjectures concerning three-valued periodic crosscorrelation functions of binary m-sequences specialize to conjectures concerning the degeneracy of quadrics of the form H^r. The main result is that if $n=2^k m$, with m odd and $k \geqq 2$, $H \subseteq PG(n-1, 2)$ is a hyperplane and H^r is a quadric, necessarily a cone of order $2l+1$, then $2l+1 \geqq 2^{k-1}+1$. This shows that when $n \equiv 0 \pmod 4$, there are no m-sequences arising from quadrics with preferred three-valued periodic crosscorrelation functions. Also, when $n=2^k$, m-sequences arising from quadrics would have three-valued periodic crosscorrelation functions with values determined by a cone of order at least $(n/2)+1$.

AMS(MOS) subject classifications. 05, 05B, 94, 94A

1. Introduction and summary. Maximum period linear recursive binary sequences of span n and period 2^n-1 possess many nice autocorrelation and crosscorrelation properties. For instance, these binary m-sequences, as they are often called, have two-valued periodic autocorrelation functions, making them useful in applications involving ranging, radar or spread-spectrum communications. In [11] the crosscorrelation properties of binary m-sequences are surveyed. For instance, it is known [3, p. 82; 5] that the periodic crosscorrelation function of two distinct m-sequences has at least three different values, and the problem of determining which pairs result in periodic crosscorrelation functions with exactly three values has received much attention (see [11] for references, also [13]). The conjectures that motivate this study are:

CONJECTURE 1 [11]. *If $n \equiv 0 \pmod 4$, then there are no pairs of binary m-sequences of span n with a preferred three-valued periodic crosscorrelation function.*

CONJECTURE 2 [5]. *If n is a power of 2, then there are no pairs of binary m-sequences of span n with a three-valued periodic crosscorrelation function.*

In this paper, a binary m-sequence of span n is viewed as a hyperplane H in $PG(n-1, 2)$—the finite projective geometry based on $GF(2^n)$, thought of as an n-dimensional vector space over $GF(2)$. Other m-sequences of span n (binary assumed throughout) then correspond to $H^r = \{P^r | P \in H\}$ for an integer r relatively prime to 2^n-1. The periodic crosscorrelation function of the sequences corresponding to H and H^r is computed by intersecting all the hyperplanes of $PG(n-1, 2)$ with H^r. A three-valued periodic crosscorrelation function results if H^r is quadric—a solution to a quadratic equation—because in this case the hyperplanes of $PG(n-1, 2)$ intersect H^r in sets of three sizes. Although not the only case when three values occur, it is the case considered here. Such quadrics are necessarily cones of order $2l$ if $n-1$ is even, and cones of order $2l+1$ if $n-1$ is odd [2]. Quadrics that are the least degenerate, i.e., $l=0$, yield pairs of m-sequences with preferred three-valued periodic crosscorrelation functions.

As special cases of Conjectures 1 and 2, we have:

CONJECTURE 1′. *If $n \equiv 0 \pmod 4$, $H \subseteq PG(n-1, 2)$ is a hyperplane, r is an integer*

* Received by the editors November 17, 1983, and in revised form August 8, 1984.

† Department of Mathematics, Colorado State University, Fort Collins, Colorado 80523. Present address. The MITRE Corporation, E-020, P.O. Box 208, Bedford, Massachusetts 01730.

relatively prime to 2^n-1, *and* H^r *is a quadric, necessarily a cone of order* $2l+1$, *then* $l \geqq 1$; *i.e.*, H^r *is a cone of order at least* 3.

CONJECTURE 2′. *If* n *is a power of* 2, $H \subseteq PG(n-1, 2)$ *is a hyperplane*, r *is an integer relatively prime to* 2^n-1, *and* H^r *is a quadric, then* H^r *is completely degenerate*; *i.e.*, H^r *is a hyperplane and* $r \equiv 2^i \bmod 2^n-1$.

The main result of this paper bears on Conjectures 1′ and 2′. In § 6 it is shown that if $n=2^k m$ with $k \geqq 2$ and m odd, $H \subseteq PG(n-1, 2)$ is a hyperplane and r is an integer relatively prime to 2^n-1; then if H^r is a quadric, it is a cone of order at least $2l+1 \geqq 2^{k-1}+1$. In particular, for $k \geqq 2$, H^r is a cone of order at least 3, and so Conjecture 1′ is true. Therefore, if there are any counterexamples to Conjecture 1, they do not arise from quadrics. When $n=2^k$, this result says a quadric H^r must be a cone of order at least $(n/2)+1$. Whether this implies that H^r must itself be a hyperplane is, along with Conjecture 2′, still open.

Actually, the results of this paper suggest a new conjecture for three-valued periodic crosscorrelation functions of m-sequences. Say that two m-sequences have a three-valued periodic crosscorrelation function of *type* l, if the values agree with the values obtained in the case of a hyperplane H and quadric H^r which is a cone of order $2l$ or $2l+1$ for $n-1$ respectively even or odd. These values do not depend on the hyperplane H. Now a preferred three-valued periodic crosscorrelation function is of type 0. Then the conjecture is:

CONJECTURE 3. *If two binary m-sequences of span* $n=2^k m$, m *odd, have a three-valued periodic crosscorrelation function of type* l, *then* $2l+1 \geqq 2^{k-1}+1$.

Conjecture 3 generalizes Conjecture 1, but is a weakening of Conjecture 2. The result in § 6 shows that Conjecture 3 is true for binary m-sequences that arise from a hyperplane $H \subseteq PG(n-1, 2)$ and a quadric H^r, r relatively prime to 2^n-1.

Section 2 shows how binary m-sequences can be viewed as hyperplanes in a finite projective geometry and gives the geometric interpretation of the shift and decimation operations. Section 3 discusses the periodic crosscorrelation function from three equivalent points of view and shows that the periodic crosscorrelation function is equivalent to hyperplane intersections in the geometry. The geometry of quadrics and the relation to pairs of m-sequences with three-valued periodic crosscorrelation functions is given in § 4. Also included in this section are results on quadrics that are needed in the proof of the main result. Section 5 contains material on linearized polynomials and subspaces of $GF(q)$ which are fixed by $x \to x^q$. These results are also needed in the proof of the main result, which is presented in § 6.

2. The geometry of sequences. By considering the points and hyperplanes of $PG(r, q)$, the finite projective geometry based on $GF(q)^{r+1}$, Singer defined a cyclic difference set $D \subseteq Z_v$, $v=(q^{r+1}-1)/(q-1)$ [4, p. 128]. If $s(D)$ is the *anti-incidence* vector of D, i.e., $s(D)_i=0$ if $i \in D$, $s(D)_i=1$ if $i \notin D$, then $s(D)$ corresponds to a binary sequence of period v with a two-valued autocorrelation function. If $q=2$ and $r=n-1$, then $s(D)$ is a binary m-sequence of span n [1, 14]. The nonzero elements of $GF(2^n)$ can be identified with the points of $PG(n-1, 2)$ and ordered using a primitive element α: $(\alpha^i | i \in Z_v, v=2^n-1)$. Then $H=\{\alpha^i \colon i \in D\}$ is a hyperplane of $PG(n-1, 2)$, and the sequence $s(D)$ is also denoted by $s(H)$. So a binary m-sequence of span n can be represented in at least these three ways: the sequence itself, as a Singer difference set in Z_v, $v=2^n-1$, and as a hyperplane in $PG(n-1, 2)$.

These connections are made more explicit with the *trace* function $\mathrm{Tr}(x)= x+x^2+\cdots+x^{2^{n-1}}$, which maps $GF(2^n)$ to $GF(2)$. Regard $GF(2^n)$ as an n-dimensional vector space over $GF(2)$. Then for fixed $\gamma \in GF(2^n)$, the mapping $x \to \mathrm{Tr}(\gamma x)$ is a linear transformation from $GF(2^n)$ to $GF(2)$. If $\gamma \neq 0$, then $H_\gamma = \{\beta \in GF(2^n) | \mathrm{Tr}\,(\gamma\beta) = 0\}$ is a subspace of $GF(2^n)$ of dimension $n-1$ (or a hyperplane of $PG(n-1, 2)$). The corresponding m-sequence is $s(H_\gamma) = (\mathrm{Tr}\,(\gamma\alpha^0), \mathrm{Tr}\,(\gamma\alpha), \cdots, \mathrm{Tr}\,(\gamma\alpha^{v-1}))$, and $\gamma \to s(H_\gamma)$ is the field isomorphism given in [7] between $GF(2^n)$ and the field formed by the $2^n - 1$ shifts of the m-sequence and the all zeros sequence.

The *shift operator* E when applied to $s = (s_0, s_1, \cdots, s_{v-1})$ yields the sequence $Es = (s_{v-1}, s_0, \cdots, s_{v-2})$. Now consider a shift of $s(H_\gamma)$:

$$\begin{aligned} Es(H_\gamma) &= (\mathrm{Tr}\,(\gamma\alpha^{-1}), \mathrm{Tr}\,(\gamma\alpha^0), \cdots, \mathrm{Tr}\,(\gamma\alpha^{v-2})) \\ &= (\mathrm{Tr}\,(\gamma\alpha^{-1}\alpha^0), \mathrm{Tr}\,(\gamma\alpha^{-1}\alpha^1), \cdots, \mathrm{Tr}\,(\gamma\alpha^{-1}\alpha^{v-1})). \end{aligned}$$

The associated hyperplane is $H_{\gamma\alpha^{-1}} = \{\beta \in GF(2^n) | \mathrm{Tr}\,(\gamma\alpha^{-1}\beta) = 0\} = \{\alpha\beta | \beta \in H_\gamma\} = \alpha H_\gamma$; i.e., $Es(H_\gamma) = s(\alpha H_\gamma)$ and the shift of the sequence corresponds to multiplication of the associated hyperplane by α. This is why a shift to the right was used in defining E. It follows that αH_γ is a hyperplane and all the hyperplanes of $PG(n-1, 2)$ are contained in $\{\alpha^i H_\gamma | i = 0, 1, 2, \cdots, v-1\}$. This is the basis of the Singer difference set construction.

Let $H \subseteq PG(n-1, 2)$ be a hyperplane and $s(H) = (s_0, s_1, \cdots, s_{v-1})$ the associated m-sequence. For an integer r consider the sequence $s(H)[r] = (t_0, t_1, \cdots, t_{v-1})$ formed by taking every r^{th} term of $s(H)$, i.e., $t_i = s_{ri}$, $i = 0, 1, \cdots, v-1$ (subscripts modulo v). The sequence $s(H)[r]$ is a *decimation* of $s(H)$ by r. It is well known that given one m-sequence of span n, then, up to cyclic shifts, all m-sequences can be obtained from it by decimating by integers r which are relatively prime to v [3, p. 78]. If $(r, v) = 1$, then the m-sequence $s(H)[r]$ equals $s(H^{r^{-1}})$, and so all m-sequences of span n, up to cyclic shifts, are obtained by considering $s(H_r)$ for a fixed hyperplane $H \subseteq PG(n-1, 2)$ and $r \in Z_v^* = \{i \in Z_v | (i, v) = 1\}$.

Example 1. Consider the primitive polynomial $f(x) = x^3 + x + 1$ over $GF(2)$. The m-sequence in this case is $s = 1001011$ which corresponds to the $(7, 3, 1)$-difference set $D = \{1, 2, 4\} \subseteq Z_7$. The nonzero elements of $GF(2^3)$ (represented in terms of α^2, α^1, α^0 coordinates) are: $\alpha^0 = 001$, $\alpha^1 = 010$, $\alpha^2 = 100$, $\alpha^3 = 011$, $\alpha^4 = 110$, $\alpha^5 = 111$, $\alpha^6 = 101$. The corresponding hyperplane is $H = \{\alpha^1, \alpha^2, \alpha^4\}$, and the other m-sequence of span 3 is obtained from $H^{-5} = H^3 = \{\alpha^3, \alpha^6, \alpha^5\}$ and is $s(H^3) = 1110100$. Figure 1 pictures $PG(2, 2)$ with the line H and the set H^3 marked.

To end this section, the above example is used to show how quadrics of $PG(n-1, 2)$ are involved in this study of m-sequences. Consider the quadratic form $Q(x_0, x_1, x_2) = x_0^2 + x_1^2 + x_2^2 + x_0x_1$. Then it can be checked that $H^3 = \{(x_0, x_1, x_2) | Q(x_0, x_1, x_2) = 0\}$, so H^3 is said to form a *quadric* (or *conic* when $n-1 = 2$) in $PG(2, 2)$. From Fig. 1 it is seen that the lines of $PG(2, 2)$ intersect H^3 in sets of size 0, 1, and 2. This is implied by a theorem given in § 4 and will be seen to be equivalent to the fact that the m-sequences $s(H)$ and $s(H^3)$ have a three-valued periodic crosscorrelation function.

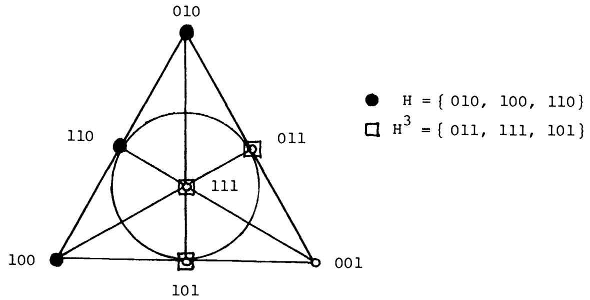


FIG. 1. *PG*(2, 2) *with H and H*^{3}.

3. Correlation of sequences. For vectors $x=(x_0, x_1, \cdots, x_{v-1})$ and $y=(y_0, y_1, \cdots, y_{v-1})$ with complex entries, the *periodic crosscorrelation function* $\theta_{x,y}$ is defined by

$$\theta_{x,y}(l)=\sum_{i=0}^{v-1} x_i y_{i-l}^*, \qquad l=0, 1, \cdots, v-1$$

where y_i^* denotes the complex conjugate of y_i and the subscripts are computed modulo v. In the case of periodic sequences over $GF(2)$, the same definition is used, but the complex conjugation is dropped and the sum is computed over the integers. In other words, if $s=(s_0, s_1, \cdots, s_{v-1})$ and $t=(t_0, t_1, \cdots, t_{v-1})$ represent two periodic binary sequences, then the periodic crosscorrelation function $\theta_{s,t}^{(1)}$ has values, for $l=0, 1, \cdots, v-1$, given by

$\theta_{s,t}^{(1)}(l)=$ number of positions that both contain 1's in the vectors s and $E^l t$.

In this paper an alternative definition of the periodic crosscorrelation function is used, namely,

$\theta_{s,t}^{(2)}(l)=$ number of positions that both contain 0's in the vectors s and $E^l t$.

Actually, in most practical applications, yet a third definition of a periodic crosscorrelation function is used. The binary sequence $s_0=(s_0, s_1, \cdots, s_{v-1})$ is replaced by the sequence $\chi(s)=(-1^{s_0}, -1^{s_1}, \cdots, -1^{s_{v-1}})$ which has entries ±1. The crosscorrelation function $\theta_{s,t}^{(3)}$ has values, for $l=0, 1, \cdots, v-1$, given by

$$\theta_{s,t}^{(3)}(l)=\theta_{\chi(s),\chi(t)}(l)=A_l-D_l$$

where

$A_l=$ number of positions that both vectors s and $E^l t$ agree

and

$D_l=$ number of positions that both vectors s and $E^l t$ disagree.

If the sequences involved all have the same *weight*, say w, i.e., the number of 1's in one period is w, and if one of the values of $\theta_{s,t}^{(1)}(l)$, $\theta_{s,t}^{(2)}(l)$ or $\theta_{s,t}^{(3)}(l)$ is known, then the values of the other two can be computed using the formulas of the next proposition, the proof of which is easy.

PROPOSITION 1. *Let $s=(s_0, s_1, \cdots, s_{v-1})$ and $t=(t_0, t_1, \cdots, t_{v-1})$ be binary vectors each of weight w and length v, and let $\theta_i=\theta_{s,t}^{(i)}(l)$, $i=1, 2, 3$. Then specifying the value of one of θ_1, θ_2, or θ_3 determines the values of the other two, namely,*

(a) *given θ_1,*

$$\theta_2=v-2w+\theta_1,$$
$$\theta_3=v-4w+4\theta_1;$$

(b) *given θ_2,*

$$\theta_1=2w-v+\theta_2,$$
$$\theta_3=4w-3v+4\theta_2;$$

(c) *given θ_3,*

$$\theta_1=(4w-v+\theta_3)/4,$$
$$\theta_2=(3v-4w+\theta_3)/4.$$

From a geometric point of view, it is best to adopt the second definition of a periodic crosscorrelation function. Suppose that $s(H)$ and $s(H^r)$ are m-sequences of span n with $H \subseteq PG(n-1,2)$ a hyperplane and r with $(r,v)=1$. Then, since the crosscorrelation counts common 0's, for $l=0,1,\cdots,v-1$,

$$\theta_{s(H^r),s(H)}(l)=|H^r \cap \alpha^l H|.$$

Now $H, \alpha H, \cdots, \alpha^{v-1}H$ represent all the hyperplanes of $PG(n-1,2)$ so that the periodic crosscorrelation function $\theta_{s(H^r),s(H)}$ is exactly the *hyperplane intersection distribution* θ_{H^r} of the set H^r.

4. The geometry of quadrics and correlations of sequences. Quadrics of the form $H^r \subseteq PG(n-1,2)$ for a hyperplane H and integer r with $(r,v)=1$ yield pairs of m-sequences $s(H)$ and $s(H^r)$ of span n with three-valued periodic crosscorrelation functions. This is based on the results of [2], [6], and [16] about the hyperplane intersection distribution of such quadrics.

In finite projective geometry $PG(N,q)$ of dimension N and order q, a prime power, the points can be taken as $(N+1)$-tuples $x=(x_0,x_1,\cdots,x_N)$ where $x_0,x_1,\cdots,x_N$ are elements of $GF(q)$ and the $(N+1)$-tuple $\rho x=(\rho x_0,\rho x_1,\cdots,\rho x_N)$ is regarded as the same point as x for any nonzero element ρ of $GF(q)$. The null $(N+1)$-tuple $(0,0,\cdots,0)$ does not represent a point. The set of points x which satisfy an equation $xC=0$ where C is a matrix of size $(N+1)\times k$ with elements in $GF(q)$ and rank k, $k=1,2,\cdots,N$, is called an *$(N-k)$-flat.*

A *quadric* Q in $PG(N,q)$ is the set of all points x which satisfy an equation $xAx^t=0$ where A is an upper triangular matrix of size $(N+1)\times(N+1)$ with elements in $GF(q)$ and x^t is the transpose of x. If m is the largest integer for which there exists a transformation of coordinates mapping Q onto a quadric Q' with the equation $xCx^t=0$, where C is an upper triangular matrix with all elements in the last m columns equal to zero, then the *rank* of Q is $N+1-m$. If $m=0$, then Q is *nondegenerate.* Otherwise, Q is called a *cone of order m* and has the form $V+Q_{N-m}$, where V is an $(m-1)$-flat called the *vertex* of Q and Q_{N-m} is a nondegenerate quadric in a complementary $(N-m)$-flat to V. Here "+" means that if $P_1 \in V$ and $P_2 \in Q_{N-m}$, then the line P_1P_2 determined by P_1 and P_2 is contained in Q. See [10] for more on quadrics.

The next theorem shows why quadrics are involved in three-valued periodic crosscorrelation functions.

THEOREM 2. *Let N be a positive integer and q a prime power. Let $Q \subseteq PG(N,q)$ be a quadric the size of a hyperplane, i.e., $|Q|=(q^N-1)/(q-1)$. Assume that Q itself is not a hyperplane.*

a. *If $N=2k$, then Q must be a cone of even order, say $2l$, and the hyperplanes of $PG(N,q)$ intersect Q in sets of three sizes with multiplicities given by*

	Size	Multiplicity
(i)	$A=\dfrac{q^{2k-1}-1}{q-1}$	$\dfrac{q^{2k+1}-1}{q-1}-q^{2k-2l}$
(ii)	$A-q^{k+l-1}$	$\dfrac{q^{k-l}(q^{k-l}-1)}{2}$
(iii)	$A+q^{k+l-1}$	$\dfrac{q^{k-l}(q^{k-l}+1)}{2}.$

b. *If $N=2k+1$, then Q must be a cone of odd order, say $2l+1$ and the hyperplanes of $PG(N,q)$ intersect Q in sets of three sizes with multiplicities given by*

	Size	Multiplicity
(i)	$B=\dfrac{q^{2k}-1}{q-1}$	$\dfrac{q^{2k+2}-1}{q-1}-q^{2k-2l}$
(ii)	$B-q^{k+l}$	$\dfrac{q^{k-l}(q^{k-l}-1)}{2}$
(iii)	$B+q^{k+l}$	$\dfrac{q^{k-l}(q^{k-l}+1)}{2}.$

Proof. See [6], [16] for the nondegenerate case when N is even and [2] for the remaining degenerate cases.

Proposition 3. *Suppose $s(H)$ and $s(H^r)$ are m-sequences of span n with $H\subseteq PG(n-1,2)$ a hyperplane and r with $(r,v)=1$. If H^r is a quadric, which is not a hyperplane, then $s(H)$ and $s(H^r)$ have a three-valued periodic crosscorrelation function. Furthermore, the ± 1 crosscorrelation values ($\theta^{(3)}$ definition earlier) and multiplicities are given by*

(i) $-1+2^{(n+e)/2}$ *occurs* $2^{n-e-1}+2^{(n-e-2)/2}$ *times,*
(ii) -1 *occurs* $2^n-2^{n-e}-1$ *times,*
(iii) $-1-2^{(n+e)/2}$ *occurs* $2^{n-e-1}-2^{(n-e-2)/2}$ *times.*

$$e=\begin{cases}2l+1, & n \text{ odd and } H^r\subseteq PG(n-1,2) \text{ is a cone of order } 2l,\\ 2l+2, & n \text{ even and } H^r\subseteq PG(n-1,2) \text{ is a cone of order } 2l+1.\end{cases}$$

Proof. Since H^r is a quadric the same size as a hyperplane, the hyperplanes of $PG(n-1,2)$ intersect H^r in the three values given in Theorem 2. Thus, the sequences $s(H)$ and $s(H^r)$ have a three-valued periodic crosscorrelation function $\theta_{s(H^r),s(H)}$. To obtain the values of $\theta^{(3)}_{s(H^r),s(H)}$ for the ± 1 case, the formulas of part (b) of Proposition 1 are used with the intersection sizes of Theorem 2 and $w=2^{n-1}$. Only one of the calculations is included for brevity.

Case 1. n is odd; $n-1=2k$ is even and $H^r\subseteq PG(n-1,2)$ is a cone of order $2l$ and $e=2l+1$. Then assuming part (aii) of Theorem 2 applies,

$$\theta_2=2^{2k-1}-1-2^{k+l-1} \text{ occurs } 2^{k-l}(2^{k-l}-1)/2 \text{ times.}$$

Then (b) of Proposition 1 gives

$$\begin{aligned}\theta_3&=2^2\cdot 2^{2k}-3(2^{2k+1}-1)+2^2(2^{2k-1}-1-2^{k+l-1})\\ &=-1-2^{k+l+1}\\ &=-1-2^{(n+2l+1)/2}\\ &=-1-2^{(n+e)/2}.\end{aligned}$$

The multiplicity becomes

$$2^{k-l}(2^{k-l}-1)/2=2^{2k-2l-1}-2^{k-l-1}=2^{n-e-1}-2^{(n-e-2)/2}. \qquad \square$$

Notice that for the values in Proposition 3, if e is large, $\theta^{(3)}_{s(H^r),s(H)}$ takes on large values, but only very few times, while if e is small, $\theta^{(3)}_{s(H^r),s(H)}$ takes on smaller values more frequently. In most instances, small values of e are desirable and in [11] a *preferred three-valued periodic crosscorrelation function* is defined to have the values given in Proposition 3 for $e=1$ if n is odd or $e=2$ if n is even. A pair of m-sequences is called a *preferred pair* if they have a preferred three-valued periodic crosscorrelation function.

For a hyperplane $H \subseteq PG(n-1,2)$ and integer r with $(r, v)=1$, if H^r is a quadric, then the sequences $s(H)$ and $s(H^r)$ are a preferred pair of m-sequences exactly when H^r is a cone of the least degeneracy. In the case that n is odd, $e=1 \Leftrightarrow 2l=0 \Leftrightarrow l=0$ and H^r is nondegenerate. In the case that n is even, $e=2 \Leftrightarrow 2l+1=1 \Leftrightarrow l=0$ and H^r is a cone of order one. In general, say that a pair of m-sequences have a *three-valued periodic crosscorrelation function of type l* if the values agree with the hyperplane intersection sizes of Theorem 2 for a cone of order $2l$, if n is odd, or $2l+1$, if n is even.

The construction of quadrics of the form H^r given in [6], [16] is related to [11, Theorem 1], and it follows that if n is odd or $n \equiv 2 \bmod 4$, there exist values of r such that $H^r \subseteq PG(n-1,2)$ is a cone of the least degeneracy. But the data of [9] shows that for $n \equiv 0 \bmod 4$ and $n \leqq 17$, there are no cones of order one of the form H^r, and indeed, if n is a power of 2 in this range, then H^r is never a quadric, except possibly the completely degenerate case of a hyperplane. Thus, Conjectures 1′ and 2′ of the introduction are obtained for quadrics of the form $H^r \subseteq PG(n-1,2)$.

This section ends with two facts about quadrics that are needed in § 6. Let $Q \subseteq PG(N, q)$ be a quadric with equation $xAx^t=0$. A point $a \in PG(N, q)$ is a *regular point* with respect to Q if $a(A+A^t) \neq 0$. Otherwise, a is called an *irregular point*. If it is clear from context that a particular quadric Q is involved, then the phrase "with respect to Q" is dropped.

THEOREM 4. *If $Q \subseteq PG(N, q)$ is a nondegenerate quadric, then if q is odd or N is odd, every point of $PG(N, q)$ is regular. When N and q are even, there is a single point of $PG(N, q)$ which is irregular. This point is called the nucleus of polarity of Q.*

THEOREM 5. *For a cone $Q \subseteq PG(N, q)$ of order m, the set of irregular points of $PG(N, q)$ consists of the points of the vertex V if either q or $N-m$ is odd. When q and $N-m$ are even, the set of irregular points of $PG(N, q)$ consists of the points of a flat $x+V$ of dimension m, where x is the nucleus of polarity of the nondegenerate quadric Q_{N-m} obtained by the intersection of Q and a $(N-m)$-flat which does not intersect the vertex V. In any case, a point of the quadric is irregular if and only if it is contained in V.*

5. Linearized polynomials and subspaces fixed by $x \rightarrow x^q$. It is well known that given a binary m-sequence s, some shift of s, say s^*, has the property that $s^*[2]=s^*$; i.e., s^* is left fixed by decimation by 2 and is called the *characteristic phase* of s. See [15] for a listing of characteristic m-sequences s^* through span $n=168$. If $H \subseteq PG(n-1,2)$ is the hyperplane that corresponds to s^*, so that $s^*=s(H)$, then $s^*[2]=s^*$ is equivalent to $H^2=H$, so H is fixed as a set by the linear map $x \rightarrow x^2$ of $GF(2^n)$. Recalling the correspondence between the shifts of an m-sequence and the elements of $GF(2^n)$, the characteristic phase of the sequence corresponds to $\gamma=1 \in GF(2^n)$. This is because the hyperplane $H_1=\{\alpha^i \in GF(2^n)\colon \operatorname{Tr}(\alpha^i)=0\}$ has $H_1^2=H_1$ since $\operatorname{Tr}(x^2)=(\operatorname{Tr}(x))^2=\operatorname{Tr}(x)$. That the linear map $\operatorname{Tr}(x)=x+x^2+\cdots+x^{2^{n-1}}$ produces such a fixed subspace of $GF(2^n)$ is a special case of a more general situation involving linearized polynomials.

A *linearized polynomial* $F(z)$ over $GF(q^n)$ is a polynomial of the form

$$F(z)=\sum_{i=0}^{h} f_i z^{q^i}, \quad f_i \in GF(q^n), \quad f_h \neq 0.$$

If the coefficients of a linearized polynomial $F(z)$ over $GF(q^n)$ in fact belong to

$GF(q)$, then $F(z)$ is called a *linearized q-polynomial.* Linearized polynomials are often used in algebraic coding theory; see, for instance, [8, Chapter 4]. One application of linearized polynomials is the characterization of the subspaces of $GF(q^n)$ that are fixed by the linear map $x \to x^q$.

THEOREM 6. *Suppose $F(z)$ is a linearized polynomial over $GF(q^n)$. The zeros of $F(z)$ form a subspace $U \subseteq GF(q^n)$ with $U^q = U$ if and only if $F(z)$ is a linearized q-polynomial.*

If $F(z) = \sum_{i=0}^{h} f_i z^{q^i}$ is a linearized q-polynomial, then $f(z) = \sum_{i=0}^{h} f_i z^i$, a polynomial over $GF(q)$, is called the *conventional associate* of $F(z)$. $F(z)$ is the *linearized associate* of $f(z)$. The subspace U of Theorem 6 is called the *fixed subspace* of F (or f).

THEOREM 7. *If F and G are linearized q-polynomials with conventional associates f and g and fixed subspaces U_f and U_g, respectively, then*

(i) *$U_f \subseteq U_G \Leftrightarrow F|G$.*

(ii) *$F|G \Leftrightarrow f|g$.*

(iii) dim *$U_f =$ degree of f.*

The next theorem bears on the crosscorrelation of $s(H^r)$ and $s(H)$.

THEOREM 8. *Suppose $n = 2^k m$, with m odd and let $H \subseteq GF(2^n)$ be the subspace of dimension $n-1$ satisfying $H^2 = H$. If s is an integer with $s|n$, i.e., $s = 2^i t$ with $0 \leqq i \leqq k$ and $t|m$, then*

(i) *The subfield $GF(2^s) \subseteq GF(2^n)$ is contained in H if and only if $i \leqq k-1$.*

(ii) *If $GF(2^s) \subseteq H$ and r is an integer with $(r, 2^n - 1) = 1$, then $GF(2^s) \subseteq H \cap H^r$.*

Proof. (i) H is the fixed subspace of the conventional associate $(x^n - 1)/(x-1) = (x-1)^{2^k - 1}((x^m - 1)/(x-1))^{2^k}$ and $GF(2^s)$ is the fixed subspace of the conventional associate $(x^s - 1) = (x-1)^{2^i}((x^t - 1)/(x-1))^{2^i}$. So, by Theorem 7, $GF(2^s) \subseteq H$ if and only if $(x^s - 1)|(x^n - 1)/(x-1)$ if and only if $i \leqq k-1$.

(ii) If $GF(2^s) \subseteq H$, then $GF(2^s)^r \subseteq H^r$. But since $(r, 2^n - 1) = 1$, $GF(2^s)^r = GF(2^s)$; $GF(2^s) \subseteq H \cap H^r$. □

COROLLARY 9. *Let $H \subseteq PG(n-1, 2)$ be a hyperplane with $H^2 = H$ and r an integer with $(r, 2^n - 1) = 1$. If $n = 2^k m$, with $k \geqq 1$ and m odd, then the crosscorrelation function $\theta_{s(H^r),s(H)}$ satisfies*

$$\theta_{s(H^r),s(H)}(0) \geqq 2^{(n/2)} - 1.$$

Proof. Let $s = 2^{k-1}m$ in the theorem. Then

$$\theta_{s(H^r),s(H)}(0) = |H^r \cap H| \geqq 2^s - 1 = 2^{(n/2)} - 1$$

since each nonzero member of $GF(2^s)$ represents a point of $PG(n-1, 2)$ which, by (ii) of Theorem 8, is contained in $H^r \cap H$. □

6. Quadrics of the form $H^r \subseteq PG(n-1, 2)$. This section contains the main result of the paper on quadrics of the form $H^r \subseteq PG(n-1, 2)$ when n is highly divisible by 2.

THEOREM 10. *Let $H \subseteq PG(n-1, 2)$ be a hyperplane, r an integer with $(r, v) = 1$ such that H^r is a quadric. If $n = 2^k m$, with $k \geqq 2$ and m odd, then H^r is a cone of order $2l+1$ with*

$$2l+1 \geqq 2^{k-1} + 1.$$

Proof. Without loss of generality, assume that $H^2 = H$. Since $N = n-1$ is odd, by Theorem 2, H^r is a cone of order $2l+1$. Let the $2l$-flat V denote the vertex of H^r and suppose Σ is a complementary $(n-2l-2)$-flat to V. Then $Q_{n-2l-2} = H^r \cap \Sigma$ is a nondegenerate quadric in Σ. Since Σ is a flat of even dimension in a projective space over a field of characteristic 2, by Theorem 4, Q_{n-2l-2} has a nucleus of polarity $x \in \Sigma$ with $x \not\in Q_{n-2l-2}$; i.e., $x \not\in H^r$. By Theorem 5, the set of irregular points with respect to H^r contained in H^r is exactly V. But $V^2 \subseteq (H^r)^2 = (H^2)^r = H^r$ and $x \to x^2$ is a nonsingular linear transformation which preserves irregularity so that necessarily $V^2 = V$. Similarly, Theorem 5 can be used to show that $(V+x)^2 = V+x$. Thus, as subspaces of $GF(2^n)$, V and $V+x$ are both fixed by $x \to x^2$, have $V \subseteq V+x$, and satisfy $\dim(V+x) = \dim(V)+1$.

Now the results of linearized polynomials are applied. There is a one-to-one correspondence between subspaces of $GF(2^n)$ fixed by $x \to x^2$ and the divisors of $z^n - 1 = z^{2^k m} - 1 = (z^m-1)^{2^k} = (z-1)^{2^k}((z^m-1)/(z-1))^{2^k}$. Suppose $f|z^n-1$ has fixed subspace V and $g|z^n-1$ has fixed subspace $V+x$, and so by Theorem 7 $f|g$, $\deg(f) = 2l+1$, and $\deg(g) = 2l+2$. If $f(x) = (z-1)^i h(z)$ with $0 \leqq i \leqq 2^k$ and $h(z)|((z^m-1)/(z-1))^{2^k}$, then since $((z^m-1)/(z-1))^{2^k}$ contains no linear factors, necessarily $g(z) = (z-1)^{i+1}h(z)$.

If $V+x \subseteq GF(2^{2^{k-1}m}) \subseteq GF(2^n)$, then by part (ii) of Theorem 8, $V+x \subseteq H \cap H^r$; i.e., $x \in H^r$, a contradiction. Thus, it must be the case that $V+x \not\subseteq GF(2^{2^{k-1}m})$. Now $GF(2^{2^{k-1}m})$ is the fixed subspace of the divisor $(z-1)^{2^{k-1}}((z^m-1)/(z-1))^{2^{k-1}}$, and by Theorem 7, $V+x \not\subseteq GF(2^{2^{k-1}m})$ if and only if $(z-1)^{i+1}h(z)$ does not divide $(z-1)^{2^{k-1}}((z^m-1)/(z-1))^{2^{k-1}}$. There are two possibilities: either $i+1 \geqq 2^{k-1}+1$ or $h(z)$ does not divide $((z^m-1)/(z-1))^{2^{k-1}}$.

Case i. If $i+1 \geqq 2^{k-1}+1$, then

$$2l+1 = \dim V = i + \deg(h(z)) \geqq i \geqq 2^{k-1},$$

so $2l+1 \geqq 2^{k-1}+1$, since $k \geqq 2$. (Only here is $k \geqq 2$ used.)

Case ii. If $h(z)$ does not divide $((z^m-1)/(z-1))^{2^{k-1}}$, then since $h(z)|((z^m-1)/(z-1))^{2^k}$, there is some factor $k(z)$ of $(z^m-1)/(z-1)$, which must have degree at least 2, such that $k(z)^{2^{k-1}+1}|h(z)$. Thus $\deg(h(z)) \geqq 2(2^{k-1}+1) = 2^k+2$ and

$$2l+1 = \dim V = i + \deg(h(z)) \geqq \deg(h(z)) \geqq 2^k+2,$$

so certainly $2l+1 \geqq 2^{k-1}+1$. □

When $k=1$ in Theorem 10, then Case i yields $2l+1 \geqq 1$, which is no constraint on l. As was mentioned previously, there are quadrics H^r in this case with $l=0$. When $k \geqq 2$, Theorem 10 shows that Conjecture 1′ of the introduction is true. More generally, Conjecture 3 of the introduction about the type of a three-valued crosscorrelation function of m-sequences is suggested.

A fact which follows from the proof of Theorem 10 that could be useful in settling Conjecture 2′, is noted.

COROLLARY 11. *Let $H \subseteq PG(n-1, 2)$ be a hyperplane, r an integer with $(r, v) = 1$ such that H^r is a quadric. If n is even, then H^r is a cone of order $2l+1$ with vertex $V \subseteq H \cap H^r$.*

Proof. In the proof of the theorem, $i+1$ must be less than 2^k; i.e., $i \leqq 2^k - 1$. Thus, the polynomial f with fixed subspace V divides $(z^n-1)/(z-1)$—the polynomial with fixed subspace H. So by theorem 7, $V \subseteq H$; i.e., $V \subseteq H \cap H^r$. □

Finally, the proof of Theorem 10 is valid for $PG(n-1, q)$ exactly when q is even, i.e., of the form $q = 2^s$. This is because the results on regular points, irregular points, and a nucleus of polarity needed in the proof hold exactly in this case. Now, though a hyperplane H with $H^q = H$ is considered, but the characterization of the fixed subspaces of $x \to x^q$ is used exactly like in the case of $q=2$. The result when q is odd is an open problem.

REFERENCES

[1] L. D. BAUMERT, *Cyclic Difference Sets*, Lecture Notes in Mathematics 182, Springer-Verlag, New York, 1971.

[2] R. A. GAMES, *The geometry of quadrics and correlations of sequences*, submitted to IEEE Trans. Inform. Theory.

[3] S. W. GOLOMB, *Shift Register Sequences*, Aegean Park Press, Laguna Hills, CA, 1982.

[4] M. HALL, *Combinatorial Theory*, John Wiley, New York, 1967.

[5] T. HELLESETH, *Some results about the cross-correlation function between two maximal linear sequences*, Discrete Math., 16 (1976), pp. 209-232.

[6] R. HOHOLDT AND J. JUSTESEN, *Ternary sequences with perfect periodic autocorrelation*, IEEE Trans. Inform. Theory, IT-29 (1983), pp. 597-599.

[7] F. J. MACWILLIAMS AND N. J. A. SLOANE, *Pseudo-random sequences and arrays*, Proc. IEEE, 64 (1976), pp. 1715-1729.

[8] ———, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.

[9] Y. NIHO, *Multi-valued cross-correlation functions between two maximal linear recursive sequences*, Ph.D. dissertation, Univ. Southern California, Los Angeles, 1972.

[10] D. K. RAY-CHAUDHURI, *Some results on quadrics in finite projective geometry based on Galois fields*, Canad. J. Math., 14 (1962), pp. 129-138.

[11] D. V. SARWATE AND M. B. PURSLEY, *Crosscorrelation properties of pseudo-random and related sequences*, Proc. IEEE, 68 (1980), pp. 593-619.

[12] D. A. SHEDD AND D. V. SARWATE, *Construction of sequences with good correlation properties*, IEEE Trans. Inform. Theory, IT-25 (1979), pp. 94-97.

[13] H. M. TRACHTENBERG, *On the cross-correlation functions of maximal linear sequences*, Ph.D. dissertation, Univ. Southern California, Los Angeles, 1970.

[14] R. J. TURYN, *On Singer's parametrization and related matters*, Appl. Res. Lab., Sylvania Electron. Syst., Waltham, MA, Eng. Note 197, 1960.

[15] M. WILLET, *Characteristic m-sequences*, Math. Comput., 30 (1976), pp. 306-311.

[16] J. WOLFMANN, *Codes projectifs à deux ou trois poids associés aux hyperquadriques d'une géométrie finie*, Discrete Math., 13 (1975), pp. 185-211.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

006

ON THE NUMBER OF REAL QUADRATIC FACTORS OF POLYNOMIALS*

ZALMAN RUBINSTEIN†

Abstract. A method for determining the number of real quadratic factors of polynomials with real or complex coefficients is introduced. The method is based on formulating an equivalent problem of finding the number of real solutions of a certain algebraic system in two variables. Criteria for the existence of special quadratic factors are also introduced. The method consists of a finite algorithm realizable on a computer. It can be applied to determine the general decomposition structure of a given real polynomial. In particular the presence or absence of real or non-real roots can be ascertained.

AMS(MOS) subject classifications. Primary 12D10; secondary 26C10, 30C15

1. Introduction. As is well known, a monic polynomial of degree n having real coefficients of the form

$$p(x)=x^n+a_1x^{n-1}+\cdots+a_n$$

can be factored as follows:

$$p(x)=(x-x_1)\cdots(x-x_k)(x-y_1)^{\alpha_1}\cdots(x-y_l)^{\alpha_l}\prod_{i=1}^{m}(x^2+c_ix+d_i)^{\beta_i} \tag{1}$$

where k $(k\geqq 0)$ is the number of real simple roots of p, l $(l\geqq 0)$ is the number of its multiple real roots and $m(m\geqq 0)$ is the number of distinct irreducible real quadratic factors of the polynomial p. Equation (1) indicates that p has k distinct simple real roots, l distinct multiple roots and m distinct pairs of conjugate complex roots. In particular, $\alpha_j\geqq 2$ for $j=1,2,\cdots,l$ and $\beta_i\geqq 0$ for $i=1,2,\cdots,m$. A lot of attention has been given to the enumeration of the various types of the linear factors of a polynomial. Most especially, there are well-known criteria for the number $(k+l)$ of distinct real roots of $p(x)$ in terms of the inner determinants or the Newton sums [4], [7], [9], [10]. In addition, the theory of discriminants yields easily the number l of real multiple roots since it is equal to the number of distinct real roots of the greatest common divisor of the polynomial p and its derivative p'. The greatest common divisor of two polynomials can be calculated by known methods [12], [13].

The purpose of this note is to indicate a method of finding the number q of distinct quadratic factors of p in (1). Obviously, since

$$q=\binom{k+l}{2}+l+m \tag{2}$$

the determination of q combined with the knowledge of k and l gives also the number m, i.e. the number of distinct conjugate pairs of zeros of p. In (2) if $k+l\leqq 1$ then $\binom{k+l}{2}$ is taken to be zero. One might mention some special cases of interest, namely when $m=0$, i.e. p has only real zeros; or when $m=[\frac{1}{2}n]$, i.e. p is of even degree and has no real zeros or p is of odd degree and has a single simple real zero. The number q, independently of m, can be thought of as a measure of simplicity of the roots of p ranging from $\binom{n}{2}$ when all the roots are real and simple to 1 when p has only one real

* Received by the editors April 24, 1984, and in revised form September 15, 1984.

† Department of Mathematics, University of Haifa, Mount Carmel, Haifa, Israel. Present address. Department of Mathematics, University of Colorado, Boulder, Colorado 80309.

multiple root, that is when $p(x)=(x-y_1)^n$ or when $p(x)=(x^2+c_1x+d_1)^{1/2n}$ in equatio (1), i.e. p has only one pair of conjugate complex roots.

The method is based on transforming the posed problem to an equivalent problei of finding the number of solutions of an algebraic system of two equations in tw independent variables. The latter problem can be treated by various methods recentl proposed in [1], [2], [3], [12], [13].

2. The main results. Consider the relation

$$(x^n+a_1x^{n-1}+\cdots+a_n)(x^2+\alpha_1x+\alpha_2)=x^{n+2}+b_1x^{n+1}+\cdots+b_{n+2}. \tag{3}$$

Equating coefficients of equal powers of x on both sides of (3) we obtain

$$\begin{aligned} &a_{-1}\alpha_2+a_0\alpha_1+a_1=b_1,\\ &a_0\alpha_2+a_1\alpha_1+a_2=b_2,\\ &\qquad\vdots\\ &a_{n-1}\alpha_2+a_n\alpha_1+a_{n+1}=b_{n+1},\\ &a_n\alpha_2+a_{n+1}\alpha_1+a_{n+2}=b_{n+2}, \end{aligned} \tag{4}$$

where we set $a_{-1}=a_{n+1}=a_{n+2}=0$, $a_0=1$ and for simplicity we assume also that $b_{n+2}\neq$ We shall need the following lemma.

LEMMA 1. *For fixed n let* $p_m=a_m\alpha_2^{n-m+1}$, $m=0,1,\cdots,n$; $p_{n+1}=p_{n+2}=0$. *Then*

$$p_m=b_{m+2}\alpha_2^{n-m}-p_{m+2}\alpha_2-p_{m+1}\alpha_1, \qquad m=0,1,\cdots,n. \tag{5}$$

Proof. For $m=n$ by (4)

$$p_n=a_n\alpha_2=b_{n+2}-p_{n+2}\alpha_2-p_{n+1}\alpha_1=b_{n+2}.$$

For $m=n-1$ one has to verify that

$$\begin{aligned} p_{n-1}=a_{n-1}\alpha_2^2&=b_{n+1}\alpha_2-p_{n+1}\alpha_2-p_n\alpha_1\\ &=b_{n+1}\alpha_2-p_n\alpha_1=b_{n+1}\alpha_2-b_{n+2}\alpha_1. \end{aligned}$$

These relations follow using the last two equations of the system (4). Assume now th (5) holds for $n-k\leqq m\leqq n$, where k is a fixed positive integer. Consider the relatio

$$\begin{aligned} A&=a_{l-1}\alpha_2+a_l\alpha_1+a_{l+1}=b_{l+1},\\ B&=a_l\alpha_2^{n-l+1}=p_l,\\ C&=a_{l+1}\alpha_2^{n-l}=p_{l+1}, \end{aligned}$$

where $l=n-k$. Now

$$D=A\alpha_2^{n-l+1}-\alpha_1B-\alpha_2C=\alpha_2^{n-l+1}b_{l+1}-p_l\alpha_1-p_{l+1}\alpha_2.$$

But also

$$\begin{aligned} D&=a_{l-1}\alpha_2^{n-l+2}+a_l\alpha_1\alpha_2^{n-l+1}+a_{l+1}\alpha_2^{n-l+1}-(a_{l+1}\alpha_2^{n-l})\alpha_2-(a_l\alpha_2^{n-l+1})\alpha_1\\ &=a_{l-1}\alpha_2^{n-l+2}=p_{l-1}. \end{aligned}$$

So (5) holds for $m=n-k-1$ and the induction is complete.

Remark. The recurrence relations (5) together with the initial conditions

$$p_n=b_{n+2}, \qquad p_{n-1}=b_{n+1}\alpha_2-b_{n+2}\alpha_1 \tag{5'}$$

determine uniquely the sequence $p_0, p_1, \cdots, p_n$ as functions of the variables α_1 and α_2 and the coefficients $b_1, b_2, \cdots, b_{n+2}$. We are ready to state the main result.

THEOREM 1. *Given a real polynomial of degree* $n+2(n \geqq 0)$

$$p(x) = x^{n+2} + b_1 x^{n+1} + \cdots + b_{n+2}$$

where $b_{n+2} \neq 0$, *the number of its distinct real quadratic factors equals the number of distinct real solutions of the algebraic system*

$$\begin{aligned} A_1(\alpha_1, \alpha_2) &= \alpha_2^{n+1} - b_2\alpha_2^n + p_1\alpha_1 + p_2\alpha_2 = 0, \\ A_2(\alpha_1, \alpha_2) &= \alpha_1\alpha_2^n - b_1\alpha_2^n + p_1 = 0, \end{aligned} \tag{6}$$

where $p_1 = p_1(\alpha_1, \alpha_2)$ *is a polynomial in* α_1 *and* α_2 *of degree at most* $n-1$ *in its variables and* $p_2 = p_2(\alpha_1, \alpha_2)$ *is a polynomial in* α_1 *and* α_2 *of degree at most* $n-2$ *in its variables. The* p_i *are obtained from the second order difference equation* (5) *with initial conditions* (5').

Proof. Put $i=0$ in (5) to obtain $p_0 = b_2\alpha_2^n - p_2\alpha_2 - p_1\alpha_1$. Also $p_0 = a_0\alpha_2^{n+1} = \alpha_2^{n+1}$. This gives the first equation of (6).

Next consider the relations $p_1 = a_1\alpha_2^n$ and $\alpha_1 + a_1 = b_1$. These imply the second equation of (6). Now it is easy to see that since $b_{n+2} \neq 0$ implies $\alpha_2 \neq 0$, each value of α_2 together with the sequence $p_1, p_2, \cdots, p_n$ determine uniquely the coefficients $a_1, \cdots, a_n$. Conversely each pair (α_1, α_2) which satisfies the system (6), where the p_i are defined by (5) and (5'), and where the a_m are defined by $a_m = p_m\alpha_2^{m-n-1}$, $m = 0, 1, \cdots, n$, satisfies the system (4). Indeed the second relation of (6) implies $\alpha_1 + a_1 = b_1$ since $\alpha_2 \neq 0$. The first relation of (6) implies, by (5) with $m=0$, that $a_0 = 1$. Hence

$$\begin{aligned} \alpha_2 + a_1\alpha_1 + a_2 &= \alpha_2 + p_1\alpha_2^{-n}\alpha_1 + p_2\alpha_2^{-n+1} \\ &= \alpha_2^{-n}(p_0 + p_1\alpha_1 + p_2\alpha_2) = b_2. \end{aligned}$$

Similarly by a straightforward induction one shows that the other equations of (4) are satisfied.

As an application of Theorem 1 consider the case $n=2$. The system (6) reduces to

$$\begin{aligned} &\alpha_2^3 - b_2\alpha_2^2 + b_4\alpha_2 + b_3\alpha_1\alpha_2 - b_4\alpha_1^2 = 0, \\ &b_1\alpha_2^2 - b_3\alpha_2 + (b_4 - \alpha_2^2)\alpha_1 = 0. \end{aligned} \tag{7}$$

Applying a standard elimination procedure [1], [2], one arrives at

$$\begin{aligned} f(\alpha_2) = \alpha_2^6 - b_2\alpha_2^5 + (b_1b_3 - b_4)\alpha_2^4 + (2b_2b_4 - b_3^2 - b_4b_1^2)\alpha_2^3 \\ + (b_1b_3b_4 - b_4^2)\alpha_2^2 - b_2b_4^2\alpha_2 + b_4^3 = 0. \end{aligned} \tag{8}$$

Equation (8) can be also derived by calculating the resultant [1] of the polynomials in (7) regarded as polynomials in the variable α_1 with coefficients depending on the parameter α_2. One assumes obviously that $\alpha_2^2 - b_4 \neq 0$.

Since in our example $\deg p = 4$, the degree of f has to be $\binom{4}{2} = 6$, the extremal case being that of p having six distinct quadratic factors and four simple real roots. In this case f has all simple real roots.

Below we indicate several examples of polynomials $p(x)$ and their corresponding $f(\alpha)$ of equation (8) and the number q defined by (2). One notices that since the second equation of (7) is linear in α_1 the number of solutions of the system (6) is equal to the number of distinct real zeros of $f(\alpha)$ provided $\alpha^2 - b_4 \neq 0$. This can be verified by the examples of Table 1.

TABLE 1

$p(x)$	$f(\alpha)$	q
$(x^2-1)(x+1)(x+2)$	$(\alpha-1)(\alpha-2)^2(\alpha+1)^2(\alpha+2)$	4
$(x^2+1)(x-1)(x+2)$	$(\alpha+1)(\alpha-1)(\alpha+i)^2(\alpha-i)^2$	2
$(x-1)^3(x+2)$	$(\alpha-1)^3(\alpha+2)^3$	2
$(x-1)^4$	$(\alpha-1)^6$	1
$(x^2+1)^2$	$(\alpha-1)^2(\alpha-i)^2(\alpha+i)^2$	1
$(x-2)(x+2)(x+1)^2$	$(\alpha-2)^2(\alpha+2)^2(\alpha-1)(\alpha+4)$	4
$(x-1)(x-2)(x-3)(x-4)$	$(\alpha-2)(\alpha-3)(\alpha-4)(\alpha-6)(\alpha-8)(\alpha-12)$	6
$(x-1)^2(x+2)^2$	$(\alpha-1)(\alpha+2)^4(\alpha-4)$	3

The factorization of $f(\alpha)$ in Table 1 is done only for illustration of Theorem 1. In general, the number of simple real zeros of $f(\alpha)$ can be found via rational operations by well-known classical means such as those mentioned in the introduction. In the general case the system (6) is a two-by-two algebraic nonlinear system. Such systems were considered recently in [2], [3]. In [1] a method of enumeration of the solutions is suggested based on successive lowering of the degree of one variable of the system. One should also note that the main purpose of these methods is to avoid the calculation of the roots of $p(z)$ which is an infinite process. Instead, the above procedure and the ones ahead are all finite algorithms realizable, for moderate n, on computers.

The question of multiple nonreducible real quadratic factors will not be treated here. Nonetheless we shall develop below a necessary criterion for the existence of such factors.

Recall that if $r(x)=r_mx^m+\cdots+r_0$ and $s(x)=s_nx^n+\cdots+s_0$, $r_ms_n\neq 0$, have a common linear factor $l(x)$, then it follows from the equations $r=lr^*$, $s=ls^*$ that

$$rs^*=sr^* \tag{9}$$

where $\deg s^*=n-1$, $\deg r^*=m-1$. If $s^*(x)=s^*_{n-1}x^{n-1}+\cdots+s^*_0$, $r^*(x)=r^*_{m-1}x^{m-1}+\cdots+r^*_0$, then equating coefficients of equal powers of x in (9) one obtains a system of $(m+n)-by-(m+n)$ homogeneous linear equations in the $(m+n)$ unknowns $s^*_0, \cdots, s^*_{n-1}$; $r^*_0, \cdots, r^*_{m-1}$ and thus one obtains the well-known condition Res. $(r,s)=0$. Now if instead of linear factors we are looking for quadratic factors $q(x)$, then (9) is satisfied with $s^*_{n-1}=r^*_{m-1}=0$, $s^*_{n-2}r^*_{m-2}\neq 0$. This time we have $(m+n-1)$ equations in $(m+n-2)$ unknowns of the above type, whose matrix

$$M=\begin{pmatrix} r_m & r_{m-1} & \cdots & r_0 & 0 & \cdots & 0 \\ 0 & r_m & r_{m-1} & \cdots & r_0 & 0\cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ 0\cdots & 0 & r_m & r_{m-1} & & \cdots & r_0 \\ s_n & s_{n-1} & \cdots & s_0 & 0 & \cdots & 0 \\ 0 & s_n & s_{n-1} & \cdots & s_0 & 0\cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & \\ 0\cdots & 0 & s_n & s_{n-1} & \cdots & s_0 & \end{pmatrix} \tag{10}$$

has $(n-1)$ rows with the r_i and $(m-1)$ rows with the s_i. This matrix is obtained from the well-known resultant matrix by deleting the last column and the nth row and $(m+n)$th row. For a nontrivial solution to exist, rank $M \leqq m+n-3$. In particular if $p(x)$ has a multiple quadratic factor the polynomials $p(x)$ and $p'(x)$ have a common quadratic factor. Consider the example $p(x)=x^5+x^3-x^2+2$, $m=5$, $n=4$, $m+n-3=6$. The seven-by-eight matrix M has the rows $(1,0,1,-1,0,2,0,0)$, $(0,1,0,1,-1,0,2,0)$, $(0,0,1,0,1,-1,0,2)$, $(5,0,3,-2,0,0,0,0)$, $(0,5,0,3,-2,0,0,0)$, $(0,0,5,0,3,-2,0,0)$, and $(0,0,0,5,0,3,-2,0)$. Its rank is seven and hence $p(x)$ has no pair of multiple roots.

Returning to the general procedure described earlier we consider the example

$$p(x)=x^5+x^3+1.$$

Here $b_1=b_3=b_4=0$, $b_2=b_5=1$, $p_3=1$, $p_2=-\alpha_1$ and $p_1=\alpha_1^2-\alpha_2$. The system (6) becomes in this case

$$\alpha_2^4-\alpha_2^3+\alpha_1^3-2\alpha_1\alpha_2=0,$$
$$\alpha_1\alpha_2^3+\alpha_1^2-\alpha_2=0.$$

This system can, by the standard elimination method mentioned before ([1], [2], [12], [13]), be brought to the form

$$\alpha_2^{10}-\alpha_2^9-2\alpha_2^5+\alpha_2^4-\alpha_2^3+1=0,$$
$$\alpha_1(\alpha_2^5-1)-\alpha_2^2=0.$$

We are led to the problem of the determination of the number of distinct real roots of the polynomial

$$z^{10}-z^9-2z^5+z^4-z^3+1=0. \tag{11}$$

The number of distinct real zeros of (11) can be found by the inners procedure [4], [9], [10]. The number of real roots is found to be two. We have $q=2$, $k=1$ and $l=0$. Thus $m=2$ by (2) and the decomposition pattern of $p(x)$ can now be determined according to formula (1). Obviously one has to apply the computer procedures available for the inner determinants to deal with the polynomial (11).

In conclusion we shall discuss a criterion for the existence of a quadratic factor of $p(x)$ of the form x^2+c. This is helpful in situations where there is importance to whether $(0,\alpha_2)$ is a possible solution of the system (6). Assume that

$$p(x)=(x^2+c)q(x)=x^n+p_1x^{n-1}+\cdots+p_n.$$

Define $p_s(x)=\frac{1}{2}(p(x)+p(-x))$, $p_a(x)=\frac{1}{2}(p(x)-p(-x))$. Then

$$p_s(x)=(x^2+c)q_s(x) \quad \text{and} \quad p_a(x)=(x^2+c)q_a(x).$$

Thus $p_s(x)$ and $p_a(x)$ have a quadratic factor in common. Conversely if $p_s(x)$ and $p_a(x)$ have a quadratic factor in common, then so does $p(x)$. These observations combined with the previous results imply the following theorem.

THEOREM 2. *For a polynomial $p(x)=x^n+p_1x^{n-1}+\cdots+p_n$ to have a quadratic factor of the form x^2+c it is necessary and sufficient that the $(n-1)$-by-$(n-1)$ matrix*

$$(12)\qquad \left(\begin{array}{cccccccc}
1 & p_2 & p_4 & \cdots & p_{2[\frac{1}{2}n]} & 0 & \cdots & 0 \\
0 & 1 & p_2 & p_4 & \cdots & p_{2[\frac{1}{2}n]} & 0 \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & & \ddots & \\
0 \cdots & 0 & 1 & p_2 & p_4 & \cdots & p_{2[\frac{1}{2}n]} & \\
p_1 & p_3 & \cdots & p_{2[\frac{1}{2}n+\frac{1}{2}]-1} & & 0 & \cdots & 0 \\
0 & p_1 & p_3 & \cdots & p_{2[\frac{1}{2}n+\frac{1}{2}]-1} & & 0 \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & & \ddots & & \\
0 \cdots & 0 & p_1 & p_3 & \cdots & p_{2[\frac{1}{2}n+\frac{1}{2}]-1} & &
\end{array}\right)$$

be singular

To illustrate condition (12) consider $p(x)=x^5+2x^3+x^2+x+1$. The corresponding singular matrix has rows $(1,2,1,0)$, $(0,1,2,1)$, $(0,1,1,0)$, $(0,0,1,1,)$. $p(x)$ has the factor x^2+1.

3. An alternate approach. The former procedure of finding q by Theorem 1 using elimination theory has one serious drawback. In some cases the elimination process introduces extraneous zeros. This is a result of the elimination being in general nonreversive. In such cases this method is not applicable to finding q through Theorem 1. The alternative of finding q by a combination of classical methods via separate calculations of k, l, and m in (2) is theoretically possible. However, this procedure is quite laborious and encounters some of the previously mentioned difficulties.

The unique advantage of the present method of Theorem 1 is that it can be used both in conjunction with elimination theory as illustrated above or with the classical method of Kronecker integrals. For a recent detailed account see [14]. This method of finding the number of simple zeros of multivariate systems is based on Gauss' theorem for multiple integrals. It gives an explicit integral representation of q (ibid. (20)-(23)) provided a domain in the (α_1, α_2)-plane containing all the zeros of the system (6) can be determined. This indeed can be done for algebraic systems. The polynomial (11) has all its roots in $|z|<3$. The possibility of this polynomial having extra roots compared to the set of α_2 in the original system is irrelevant to the final conclusion.

Now the equation $\alpha_1^2+\alpha_2^3\alpha_1-\alpha_2=0$ in the variable α_1 has all its roots in the region $|\alpha_1|<1+\max[|\alpha_2|^3, |\alpha_2|]$. Since $|\alpha_2|<3$ we have $|\alpha_1|<28$ so that all the zeros of our system lie in the rectangle R in the (α_1, α_2)-plane defined by $|\alpha_1|<28$ and $|\alpha_2|<3$. The calculation of q can be done by evaluating the Kronecker integral with respect to R. Obviously since the result of integration is an integer, the problem of accuracy is mild when using approximate calculations.

4. Conclusions. It is clear from the method and examples indicated above that the price of substituting an infinite procedure of finding all the roots of a polynomial by a finite procedure to solve the problem of determining the number of quadratic factors of a real polynomial is the relatively large order of the polynomial obtained at the end of the elimination procedure. In addition, procedures for manipulation of multivariable polynomials are required. Fortunately such procedures have been worked out, e.g. [12], [13].

One might add in passing that although this discussion has been centered around polynomials with real coefficients there is no difficulty to treating polynomials with complex coefficients. In this case the system (6) is an algebraic system with complex coefficients. It is easy to extend the method of enumerating the distinct real zeros of real polynomials to complex polynomials. Indeed if $p(z)=p_1(z)+ip_2(z)$, where $p_1(z)$ and $p_2(z)$ are real polynomials, every real zero of $p(z)$ is a real zero of both $p_1(z)$ and $p_2(z)$ and hence a real zero of the greatest common divisor of $p_1(z)$ and $p_2(z)$ and vice versa. Thus the problem of finding the number of distinct real roots of the complex polynomial $p(z)$ is equivalent to finding the number of the distinct real roots of an easily computable real polynomial.

Enumeration problems for polynomials have a definite applied character (see for example the introduction of [14]; [1], [2], [5]). Recently progress has also been made to tackle some of the more difficult theoretical questions such as the problem posed by S. Karlin to characterize all the finite zero diminishing transformations on a given polynomial [11], [6].

Acknowledgment. The author wishes to thank Mr. Simcha Brudno for some valuable conversations and several suggestions relating to this note.

REFERENCES

[1] A. Benallou, D. A. Mellichamp and D. E. Seborg, *Characterization of equilibrium sets for bilinear systems with feedback control*, Automatika, 19 (1983), pp. 183-189.

[2] ———, *On the number of solutions of multivariate polynomial systems*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 224-227.

[3] N. K. Bose, *Test for Lyapunov stability by rational operations*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 700-702.

[4] N. K. Bose and E. I. Jury, *Inner algorithm to test for positive definiteness of arbitrary binary forms*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 169-170.

[5] N. K. Bose, *Multidimensional Systems Theory and Applications*, IEEE Press, New York, 1979.

[6] T. Craven and G. Csordas, *Zero diminishing linear transformations*, Proc. Amer. Math. Soc., 80 (1980), 544-546.

[7] F. R. Gantmacher, *Theory of Matrices, Vol.* II, Chelsea, New York, 1959, pp. 201-204.

[8] C. B. Garcia and T. Y. Li, *On the number of solutions to polynomial systems of equations*, SIAM J. Numer. Anal., 17 (1980), pp. 540-546.

[9] E. I. Jury and S. M. Ahn, *A computational algorithm for inners*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 541-543.

[10] P. S. Kamat, *On computational algorithm for inners*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 154-155.

[11] S. Karlin, *Total Positivity, Vol.* 1, Stanford University Press, Stanford, CA, 1968, pp. 381-382.

[12] J. Moses, *Solution of systems of polynomial equations by elimination*, Comm. ACM., 9 (1966), pp. 634-637.

[13] H. L. Williams, *Algebra of polynomials in several variables for a digital computer*, J. ACM, 9 (1962), pp. 29-40.

[14] C. H. Stump and B. J. Hoenders, *On the calculation of the exact number of zeros of a set of equations*, Computing, 30 (1983), pp. 137-147.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

007

LABELLED GRAPHS WITH SMALL VERTEX DEGREES AND *P*-RECURSIVENESS*

I. P. GOULDEN† AND D. M. JACKSON†

Abstract. We show that the number of labelled graphs with vertices of degrees 1, 2, 3 or 4 only satisfy linear recurrence equations, and are therefore *P*-recursive. We conjecture that the number of labelled graphs with vertices whose degrees belong to a given finite set is also *P*-recursive.

AMS(MOS) subject classifications. 05C30, 05A15

1. Introduction. A sequence $\{a_n | n \geqq 0\}$ is said to be *P-recursive* if it satisfies a homogeneous linear recurrence equation of finite order, with polynomial coefficients. Such sequences are of interest because the n-th term can be computed in time that is linear in n, and space that is independent of n. The formal power series $A(x) = \sum_{n\geqq 0} a_n x^n/n!$, called the exponential generating function for $\{a_n | n \geqq 0\}$, is said to be *D-finite* if A satisfies a linear homogeneous differential equation of finite order, whose coefficients are polynomials in x. Stanley [8] discusses the equivalence of the D-finiteness of A and the P-recursiveness of $\{a_n | n \geqq 0\}$, as well as showing that many combinatorially defined power series are D-finite.

For $\alpha \subset \{0, 1, \cdots\}$, let $\mathrm{G}_{0,\alpha}$ be the set of labelled graphs, each of whose vertex degrees lies in α, and let $\mathrm{G}_{1,\alpha}$ denote the set of simple graphs in $\mathrm{G}_{0,\alpha}$. Suppose that the number of graphs on n vertices in $\mathrm{G}_{i,\alpha}$ is denoted by $g_{i,\alpha}(n)$, and that the exponential generating function for $\mathrm{G}_{i,\alpha}$ with respect to vertices is $G_{i,\alpha}(x) = \sum_{n\geqq 0} g_{i,\alpha}(n)x^n/n!$, for $i = 0, 1$. A p-regular graph is one in which each vertex has degree p, and corresponds to the choice $\alpha = \{p\}$ above.

Read [5] has shown that $G_{1,\{3\}}$ is D-finite, and it is implicit in Read and Wormald [6] that $G_{1,\{4\}}$ is D-finite. Goulden, Jackson and Reilly [2] have shown that $G_{0,\{3\}}$ and $G_{0,\{4\}}$ are D-finite. Stanley [8] has asked whether $G_{i,\{p\}}$ is D-finite for all p. In this paper we consider sets α of vertex-degrees with more than a single element. Applying the methods developed in Goulden, Jackson and Reilly [2], we construct differential equations which demonstrate that $G_{i,\alpha}$ is D-finite for $i = 0, 1$ and all choices of α whose maximum element (denoted by $m(\alpha)$) is less than or equal to 4.

Throughout this paper we denote the coefficient of $x_1^{i_1}x_2^{i_2}\cdots$ in the formal power series $f(x_1, x_2, \cdots)$ by $[x_1^{i_1}x_2^{i_2}\cdots]f$. For details of the sum and product lemmas for labelled configurations see Goulden and Jackson [1].

2. Preliminary cases. Certain $G_{i,\alpha}$ can be obtained immediately by elementary combinatorial arguments, using only the sum and product lemmas for exponential generating functions. The first simplification is to note that $G_{i,\{0\}\cup\alpha} = e^x G_{i,\alpha}$, for $0 \notin \alpha$, $i = 0, 1$. Thus $G_{i,\{0\}\cup\alpha}$ is D-finite if and only if $G_{i,\alpha}$ is D-finite, and so it is enough to consider only the case $\alpha \subset \{1, 2, \cdots\}$ in the remainder of this paper.

For the case $m(\alpha) = 1$, we immediately have $G_{i,\{1\}} = \exp(x^2/2)$ for $i = 0, 1$ since, for labelled graphs with only vertices of degree 1, the connected components are single edges, each of which has generating function $x^2/2$.

* Received by the editors August 25, 1983, and in revised form September 17, 1984. This work was supported by the Natural Sciences and Engineering Research Council of Canada under grants U0073 and A8235.

† Faculty of Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

For the case $m(\alpha)=2$, we consider labelled graphs whose connected components are paths or cycles. Thus

$$G_{0,\{2\}}=(1-x)^{-1/2}\exp\left(\frac{x}{2}+\frac{x^2}{4}\right),\qquad G_{1,\{2\}}=(1-x)^{-1/2}\exp\left(-\frac{x}{2}-\frac{x^2}{4}\right),$$

$$G_{0,\{1,2\}}=(1-x)^{-1/2}\exp\left(\frac{x}{2}+\frac{x^2}{4}+\frac{x^2}{2(1-x)}\right),$$

$$G_{1,\{1,2\}}=(1-x)^{-1/2}\exp\left(-\frac{x}{2}-\frac{x^2}{4}+\frac{x^2}{2(1-x)}\right),$$

so for $m(\alpha)\leqq 2$ and $i=0,1$, we have directly obtained an expression for $G_{i,\alpha}$. Differentiating these expressions once, we immediately obtain the first order differential equation $\phi_1(d/dx)G_{i,\alpha}+\phi_0 G_{i,\alpha}=0$, where ϕ_1 and ϕ_0 are given explicitly for each such i and α in Table 1.

TABLE 1
Differential equations for $G_{i,\alpha}(x)$ with $m(\alpha)\leqq 2$.

α	i	ϕ_1	ϕ_0
$\{1\}$	0	1	$-x$
$\{1\}$	1	1	$-x$
$\{2\}$	0	$2(1-x)$	x^2-2
$\{2\}$	1	$2(1-x)$	$-x^2$
$\{1,2\}$	0	$2(1-x)^2$	$-x^3+2x^2-2$
$\{1,2\}$	1	$2(1-x)^2$	$x(x^2-2)$

For the cases $m(\alpha)=3$ and $m(\alpha)=4$, we have no explicit expression for $G_{i,\alpha}(x)$, so we cannot proceed as we have in the previous cases $m(\alpha)=1,2$. Instead, we follow the indirect procedure given in the next section.

3. Symmetric multivariate generating functions for $m(\alpha)=3$, 4. Suppose that we are interested in the sequence $\{c_p(n)\mid n\geqq 0\}$ where $c_p(n)=[t_1^p\cdots t_n^p]T(\mathbf{t})$, and $T(\mathbf{t})$ is a symmetric function in the indeterminates $\mathbf{t}=(t_1,t_2,\cdots)$. We say that $c_p(n)$ is a *regular coefficient* of $T(\mathbf{t})$. Further suppose that $T(\mathbf{t})$ is expressed in terms of the power sum symmetric functions $s_i=\sum_{j\geqq 1}t_j^i$ as $T(\mathbf{t})=E(\mathbf{s})$, where $\mathbf{s}=(s_1,s_2,\cdots)$. Then $c_p(n)=[y_p^n/n!]V(y_1,\cdots,y_p)$, by the H-series theorem (Goulden, Jackson and Reilly [2]) where $V(=H(E)$, the *H-series* of $E)$ is the solution to a system of p partial differential equations derived from a system of partial differential equations for E itself. If these equations for V can be manipulated in a way that eliminates all differentiation with respect to $y_1,\cdots,y_{p-1}$, we can then set $y_1=\cdots=y_{p-1}=0$ to obtain an ordinary differential equation for $V(0,\cdots,0,y_p)=\sum_{n\geqq 0}c_p(n)y_p^n/n!$, and hence deduce the D-finiteness of $V(0,\cdots,0,y_p)$. This procedure has been followed for 3- and 4-regular graphs in [2]. The following result enables us to carry it out for sets α with more than a single element.

PROPOSITION 3.1.

$$g_{i,\alpha}(n)=[t_1^{m(\alpha)}\cdots t_n^{m(\alpha)}]\prod_{j\geqq 1}\left(\sum_{k\in\alpha}t_j^{m(\alpha)-k}\right)T_i\quad \textit{for } i=0,1$$

where

$$T_0=\prod_{1\leqq l\leqq j}(1-t_lt_j)^{-1},\qquad T_1=\prod_{1\leqq l<j}(1+t_lt_j).$$

Proof. $[t_1^{d_1}\cdots t_n^{d_n}]T_i$ is the number of labelled graphs in which the vertex with label k has degree d_k, for $k=1,\cdots,n$, when $i=0$. In the case $i=1$, we have the number of such graphs that are simple. Thus

$$\begin{aligned} g_{i,\alpha}(n) &= \sum_{d_1\in\alpha}\cdots\sum_{d_n\in\alpha}[t_1^{d_1}\cdots t_n^{d_n}]T_i \\ &= \sum_{d_1\in\alpha}\cdots\sum_{d_n\in\alpha}[t_1^{m(\alpha)}\cdots t_n^{m(\alpha)}]t_1^{m(\alpha)-d_1}\cdots t_n^{m(\alpha)-d_n}T_i \\ &= [t_1^{m(\alpha)}\cdots t_n^{m(\alpha)}]\prod_{j=1}^{n}\left(\sum_{k\in\alpha} t_j^{m(\alpha)-k}\right)T_i \end{aligned}$$

and the result follows, since $(\sum_{k\in\alpha} t_j^{m(\alpha)-k})|_{t_j=0}=1$. □

This result gives the required numbers of graphs as regular coefficients in symmetric power series. For each i and α, with $m(\alpha)=3$ or 4, we denote the expression for this symmetric power series in terms of $\mathbf{s}$ by $E_{i,\alpha}(\mathbf{s})$ and determine $E_{i,\alpha}(\mathbf{s})$ by applying exp log to the generating function in Proposition 3.1. For example,

$$\begin{aligned} g_{0,\{1,2,4\}}(n) &= [t_1^4\cdots t_n^4]\prod_{j\geqq 1}(1+t_j^2+t_j^3)T_0 \\ &= [t_1^4\cdots t_n^4]\prod_{j\geqq 1}(1+t_j^2)(1-t_j^3)^{-1}T_0 \\ &= [t_1^4\cdots t_n^4]\exp\left\{\sum_{j\geqq 1}\log(1+t_j^2)+\log(1-t_j^3)^{-1}+\sum_{l\leqq j}\log(1-t_lt_j)^{-1}\right\} \\ &= [t_1^4\cdots t_n^4]\exp\left\{\sum_{j\geqq 1}\sum_{k\geqq 1}\frac{1}{k}((-1)^{k-1}t_j^{2k}+t_j^{3k})+\sum_{l\leqq j}\sum_{k\geqq 1}\frac{1}{k}t_l^kt_j^k\right\}, \end{aligned}$$

so that

$$E_{0,\{1,2,4\}}(\mathbf{s})=\exp\left\{\sum_{k\geqq 1}\frac{1}{k}(s_{3k}+(-1)^{k-1}s_{2k}+(s_k^2+s_{2k})/2)\right\}.$$

Similarly, for all α with $m(\alpha)=3$ or 4, $E_{i,\alpha}(\mathbf{s})=\exp\{a_i+b_\alpha\}$, where

$$a_0=\sum_{k\geqq 1}(s_k^2+s_{2k})/2k,\qquad a_1=\sum_{k\geqq 1}(-1)^{k-1}(s_k^2-s_{2k})/2k$$

and the b_α, for $m(\alpha)=3$ or 4, are given in Table 2.

TABLE 2

Power sum representations for $\log(G_{i,\alpha})-a_i$ *with* $m(\alpha)=3, 4$.

α	b_α
$\{3\}$	1
$\{1,3\}$	$\sum_{k\geqq 1}s_{2k}/k$
$\{2,3\}$	$\sum_{k\geqq 1}(-1)^{k-1}s_k/k$
$\{1,2,3\}$	$\sum_{k\geqq 1}(s_k-s_{3k})/k$
$\{4\}$	1
$\{1,4\}$	$\sum_{k\geqq 1}s_{3k}/k$
$\{2,4\}$	$\sum_{k\geqq 1}(-1)^{k-1}s_{2k}/k$
$\{3,4\}$	$\sum_{k\geqq 1}(-1)^{k-1}s_k/k$
$\{1,2,4\}$	$\sum_{k\geqq 1}(s_{3k}+(-1)^{k-1}s_{2k})/k$
$\{1,3,4\}$	$\sum_{k\geqq 1}(s_{3k}-s_{4k}+(-1)^{k-1}s_k)/k$
$\{2,3,4\}$	$\sum_{k\geqq 1}(s_k-s_{3k})/k$
$\{1,2,3,4\}$	$\sum_{k\geqq 1}(s_k-s_{4k})/k$

Of course, $g_{i,\alpha}(n)=[t_1^3\cdots t_n^3]E_{i,\alpha}(\mathbf{s})$ for $m(\alpha)=3$, and $g_{i,\alpha}(n)=[t_1^4\cdots t_n^4]E_{i,\alpha}(\mathbf{s})$ for $m(\alpha)=4$.

4. Univariate generating functions for $m(\alpha)=3$, 4. It is now a straightforward matter to obtain a system of partial differential equations for $E_{i,\alpha}(\mathbf{s})$. For example

$$k\frac{\partial}{\partial s_k}E_{0,\{1,2,4\}}=\begin{cases}(4-2(-1)^{k/2}+s_k)E_{0,\{1,2,4\}}, & k=0\ (\mathrm{mod}\ 6)\\ (3+s_k)E_{0,\{1,2,4\}}, & k=3\ (\mathrm{mod}\ 6)\\ (1-2(-1)^{k/2}+s_k)E_{0,\{1,2,4\}}, & k=2,4\ (\mathrm{mod}\ 6)\\ s_kE_{0,\{1,2,4\}}, & k=1,5\ (\mathrm{mod}\ 6).\end{cases}$$

Carrying this out for all α with $m(\alpha)=3$, we find that the H-series $V(y_1,y_2,y_3)=H(E_{i,\alpha})$ satisfies the system

$$\begin{aligned}&V_1=(c+y_1)V+y_2V_1+y_3V_2,\\ &2V_2-V_{11}=(d+fy_2)V+fy_3V_1,\\ &3V_3-3V_{12}+V_{111}=(e+y_3)V,\end{aligned}\tag{1}$$

where $V_{ij\cdots}$ denotes $\partial/\partial y_i\,\partial/\partial y_j\cdots V$, and the values of c, d, e, f corresponding to each (i,α) are given in Table 3.

TABLE 3
Parameter values for system (1).

α	c	d	e	i	f
$\{3\}$	0	f	0	0	1
$\{1,3\}$	0	$2+f$	0	1	-1
$\{2,3\}$	1	$-1+f$	1		
$\{1,2,3\}$	1	$1+f$	2		

For $m(\alpha)=4$, the H-series $V(y_1,y_2,y_3,y_4)$ satisfies the system

$$\begin{aligned}&V_1=(c+y_1)V+y_2V_1+y_3V_2+y_4V_3,\\ &2V_2-V_{11}=(d+gy_2)V+gy_3V_1+gy_4V_2,\\ &3V_3-3V_{12}+V_{111}=(e+y_3)V+y_4V_1,\\ &4V_4-4V_{13}-2V_{22}+4V_{112}-V_{1111}=(f+gy_4)V,\end{aligned}\tag{2}$$

where the values of c, d, e, f, g corresponding to each (i,α) are given in Table 4.

TABLE 4
Parameter values for system (2).

α	c	d	e	f	i	g
$\{4\}$	0	g	0	1	0	1
$\{1,4\}$	0	g	3	1	1	-1
$\{2,4\}$	0	$2+g$	0	-1		
$\{3,4\}$	1	$-1+g$	1	0		
$\{1,2,4\}$	0	$2+g$	3	-1		
$\{1,3,4\}$	1	$-1+g$	4	-4		
$\{2,3,4\}$	1	$1+g$	-2	2		
$\{1,2,3,4\}$	1	$1+g$	1	-2		

The two special cases of system (1) corresponding to 3-regular graphs and simple graphs have been given in [2]. If we remove all partial derivatives with respect to y_1 and y_2 from system (1) by means of the elimination scheme given in [2], and then set $y_1 = y_2 = 0$, we obtain a second order differential equation for $G_{i,\alpha}(x) = V(0, 0, x)$. If this equation is denoted by

$$\phi_2(x)\frac{d^2}{dx^2}G_{i,\alpha}(x) + \phi_1(x)\frac{d}{dx}G_{i,\alpha}(x) + \phi_0(x)G_{i,\alpha}(x) = 0,$$

then the values of ϕ_0, ϕ_1, ϕ_2 for each (i, α) with $m(\alpha) = 3$ are given in Table A of the Appendix. The values of $g_{i,\alpha}(n)$, $n = 0, \cdots, 10$, deduced from the differential equations are given in Table B, for checking purposes.

Similarly, two special cases of system (2) have been given in [2]. The elimination scheme which was used in [2] to obtain a second order differential equation for $G_{i,\alpha}(x) = V(0, 0, 0, x)$ will only work in 4 of the 16 cases that arise from $m(\alpha) = 4$ (including the two cases reported in [2]). This is because our elimination scheme involved finding linear equations in derivatives with respect to y_1 and y_4. For 4 sets of values of c, d, e, f, g, the two equations given in [2] involve only V_{44}, V_4, V, V_{11}, so V_{11} is eliminated to yield a second order ordinary differential equation. For the other 12 sets of parameter values, the two equations involve V_{44}, V_4, V, V_{11}, V_1. Thus we derive a third equation from these, involving V_{444}, V_{44}, V_4, V, V_{11}, V_1, and eliminate V_{11}, V_1 between these three equations to yield a third order differential equation.

Since these third order differential equations have large polynomials as coefficients, we do not give them here. The four cases with second order differential equations are $i = 0$, 1 and $\alpha = \{4\}$, $\{2, 4\}$. The cases with $\alpha = \{4\}$ have been reported in [2], so we omit them, and give the values of ϕ_0, ϕ_1, ϕ_2, for the differential equation

$$\phi_2(x)\frac{d^2}{dx^2}G_{i,\alpha}(x) + \phi_1(x)\frac{d}{dx}G_{i,\alpha}(x) + \phi_0(x)G_{i,\alpha}(x) = 0$$

with $\alpha = \{2, 4\}$ in Table C of the Appendix. The values of $g_{i,\{2,4\}}(n)$ for $n = 0, \cdots, 10$ are given in Table D.

5. A conjecture. In general, for any α, it is routine to derive a system of $m(\alpha)$ partial differential equations for $V(y_1, y_2, \cdots, y_{m(\alpha)})$. These can, of course, be transformed into a system of simultaneous recurrence equations in $m(\alpha)$ dimensions, which can be used to give the required number, $g_{i,\alpha}(n) = [y^n_{m(\alpha)}/n!]V$, in time which is of order $n^{m(\alpha)}$. To enable us to calculate $g_{i,\alpha}(n)$ in time which is linear in n, we must first reduce the system of partial differential equations for $V(y_1, \cdots, y_{m(\alpha)})$ to a single ordinary differential equation for $V(0, \cdots 0, y_{m(\alpha)})$, as we have done in the previous section when $m(\alpha) = 3, 4$. When $m(\alpha) \geqq 5$, we can find elimination schemes to perform this reduction, but the computation becomes very lengthy. For example, for the 5-regular simple graphs, with $i = 1$, $\alpha = \{5\}$, we have carried out the very time-consuming elimination, and have obtained a differential equation for $G_{1,\{5\}}(x)$. Unfortunately, it is of sixth order, and the degrees of the polynomial coefficients exceed 100. The first 20 values of $g_{1,\{5\}}(n)$, deduced from this equation, agree with the results of McKay [4]. This differential equation demonstrates that $G_{1,\{5\}}(x)$ is D-finite, but there is certainly no guarantee that it is the lowest-order ordinary differential equation with polynomial coefficients which can be found for $G_{1,\{5\}}(x)$.

The differential equations that we have obtained lead us to make the following conjecture.

CONJECTURE 5.1. *The numbers $g_{0,\alpha}(n)$ and $g_{1,\alpha}(n)$, of labelled graphs and simple labelled graphs, respectively, with n vertices, each with degree in α, are P-recursive for any finite α.*

From the results of this paper, it seems that k-regular graphs are computationally equivalent to graphs whose vertex-degrees lie in α, where α has maximum element k. It might be that certain choices of α, say $\alpha=\{0,1,\cdots,k\}$ would be more convenient to work with, in proving P-recursiveness, than k-regular graphs because of more "freedom" in constructions, while yielding equivalent results.

6. Plane partitions. If $p(i_1,\cdots,i_n)$ is the number of plane partitions with i_j copies of j for $j=1,\cdots,n$, then

$$\begin{aligned} p(i_1,\cdots,i_n)&=[t_1^{i_1}\cdots t_n^{i_n}]\prod_{j\geqq 1}(1-t_j)^{-1}\prod_{l<j}(1-t_lt_j)^{-1}\\ &=[t_1^{i_1}\cdots t_n^{i_n}]\prod_{j\geqq 1}(1+t_j)\prod_{l\leqq j}(1-t_lt_j)^{-1}, \end{aligned}$$

from Stanley [7] or Macdonald [3]. Thus if $q_m(n)$ is the number of plane partitions with m copies of each of $1,2,\cdots,n$, then

$$q_m(n)=g_{0,\{m-1,m\}}(n).$$

Thus, we have demonstrated that $\{q_m(n)\,|\,n\geqq 0\}$ is P-recursive for $m\leqq 4$, and conjecture that it is P-recursive for all m.

Appendix.

TABLE A

Polynomial coefficients in ordinary differential equations for $G_{i,\alpha}(x)$ when $m(\alpha)=3$.

i	α	j	ϕ_j
0	$\{3\}$	0	$x(x^{10}-10x^8+24x^6-4x^4-44x^2-48)$
		1	$-3(x^{10}-6x^8+9x^6+18x^4+10x^2-8)$
		2	$9x^3(x^4-2x^2-2)$
0	$\{1,3\}$	0	$x(x^{10}-18x^8+120x^6-272x^4-324x^2-120)$
		1	$-3(x^{10}-14x^8+41x^6+36x^4+2x^2-8)$
		2	$9x^3(x^4-4x^2-2)$
0	$\{2,3\}$	0	$x^{11}+x^{10}-6x^9-4x^8+11x^7-15x^6+8x^5-2x^3+12x^2-24x-24$
		1	$-3(x^{10}-2x^8+2x^6-6x^5+8x^4+2x^3+8x^2+16x-8)$
		2	$9x^3(x^4-x^2+x-2)$
0	$\{1,2,3\}$	0	$x^{11}-2x^{10}-14x^9+24x^8+74x^7-61x^6-99x^5$ $-55x^4-180x^3-48x^2-96x-24$
		1	$-3(x^{10}-10x^8-6x^7+22x^6+8x^5+20x^4+26x^3+16x-8)$
		2	$9x^3(x+2)(x^3-2x^2+x-1)$
1	$\{3\}$	0	$-x^3(x^4+2x^2-2)^2$
		1	$3(x^{10}+6x^8+3x^6-6x^4-26x^2+8)$
		2	$9x^3(x^4+2x^2-2)$
1	$\{1,3\}$	0	$-x(x^4-4x^2+2)(x^6-2x^2+12)$
		1	$3(x^{10}-2x^8-5x^6-18x^2+8)$
		2	$9x^3(x^4-2)$
1	$\{2,3\}$	0	$-x^2(x^9+x^8+8x^7+14x^6+15x^5+9x^4-24x^3-22x^2+16x+12)$
		1	$3(x^{10}+10x^8-4x^7+16x^6-2x^5-14x^4+34x^3-24x^2-16x+8)$
		2	$9x^3(x^4+3x^2+x-2)$
1	$\{1,2,3\}$	0	$-x(x^{10}-2x^9-6x^7-12x^6+x^5-x^4+39x^3-10x^2+24)$
		1	$3(x^{10}+2x^8+2x^7-4x^6+8x^5-2x^4+10x^3-16x^2-16x+8)$
		2	$9x^3(x^4+x^2+x-2)$

TABLE B
Initial values for $g_{i,\alpha}(n)$ when $m(\alpha)=3$.

i	α	$\{g_{i,\alpha}(n) \mid 0 \leqq n \leqq 10\}$
0	$\{3\}$	1, 0, 2, 0, 47, 0, 4720, 0, 1256395, 0, 699971370
0	$\{1, 3\}$	1, 0, 5, 0, 186, 0, 22960, 0, 6831650, 0, 4071581010
0	$\{2, 3\}$	1, 1, 4, 23, 214, 2698, 44288, 902962, 22262244, 68446612, 21940389584
0	$\{1, 2, 3\}$	1, 1, 7, 47, 521, 7233, 129443, 2811701, 73203561, 2229207953, 78389689559
1	$\{3\}$	1, 0, 0, 0, 1, 0, 70, 0, 19355, 0, 11180820
1	$\{1, 3\}$	1, 0, 1, 0, 8, 0, 730, 0, 188790, 0, 102737670
1	$\{2, 3\}$	1, 0, 0, 1, 10, 112, 1760, 35150, 848932, 24243520, 805036704
1	$\{1, 2, 3\}$	1, 0, 1, 4, 41, 512, 8285, 166582, 4054953, 116797432, 3912076929

TABLE C
Polynomial coefficients in ordinary differential equations for $G_{i,\{2,4\}}(x)$, $i=0, 1$.

i	j	ϕ_j
0	0	$(-x^{14}+6x^{13}+2x^{12}-76^{11}+112x^{10}+96x^9+356x^8-1320x^7$ $-568x^6+768x^5+9248x^4+12224x^3-2496x^2-3968x-768)$
	1	$4(x^{13}-4x^{12}-6x^{11}+36x^{10}-6x^9+24x^8-352x^7+380x^6$ $+152x^5+2104x^4-1472x^3-688x^2+256x+96)$
	2	$-16(x-2)^2x^2(x+1)^2(x^5-2x^4+2x^3-2x^2+12x+4)$
1	0	$x^2(x^{12}+6x^{11}+14x^{10}+12x^9-16x^8+24x^7+116x^6-184x^5$ $-456x^4+480x^3+512x^2-704x+192)$
	1	$4(x^{13}+4x^{12}-2x^{11}-20x^{10}+2x^9+40x^8-104x^7-204x^6$ $+200x^5+328x^4-288x^3-208x^2+320x-96)$
	2	$-16(x-1)^2x^2(x+2)^2(x^2+2x-2)(x^3+2)$

TABLE D
Initial values for $g_{0,\{2,4\}}(n)$ and $g_{1,\{2,4\}}(n)$.

i	$\{g_{i,\{2,4\}}(n) \mid 0 \leqq n \leqq 10\}$
0	1, 2, 9, 65, 751, 13044, 320803, 10609256, 453774440, 24375801464, 1607240682376
1	1, 0, 0, 1, 3, 38, 730, 20670, 781578, 37885204, 2289786624

Acknowledgments. The calculations were carried out by the symbolic algebra system called VAXIMA at the University of Waterloo. VAXIMA is based on the MACSYMA system developed at the Massachusetts Institute of Technology.

REFERENCES

[1] I. P. GOULDEN AND D. M. JACKSON, *Combinatorial Enumeration*, John Wiley, New York, 1983.

[2] I. P. GOULDEN, D. M. JACKSON AND J. W. REILLY, *The Hammond series of a symmetric function and its application to P-recursiveness*, this Journal, 4 (1983), 179–193.

[3] I. G. MACDONALD, *Symmetric Functions and Hall Polynomials*, Oxford Univ. Press, New York, 1979.

[4] B. D. MCKAY, *Applications of a technique for labelled enumeration*, preprint.

[5] R. C. READ, *The enumeration of locally restricted graphs* (II), J. Lond. Math. Soc., 35 (1960), pp. 334–351.

[6] R. C. READ AND N. C. WORMALD, *Number of labelled 4-regular graphs*, J. Graph Theory, 4 (1980), pp. 203–212.

[7] R. P. STANLEY, *Theory and applications of plane partitions: Parts* I, II, Studies Appl. Math., 50 (1971), pp. 167–188, pp. 259–279.

[8] ———, *Differentiably finite power series*, European J. Comb., 1 (1980), pp. 175–188.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

008

AN ANALOGUE OF THE SHANNON CAPACITY OF A GRAPH*

MARTIN FARBER†

Abstract. The Shannon capacity of a graph G is the value $\alpha_s(G)=\sup_n \sqrt[n]{\alpha(G^n)}$, where $\alpha(G^n)$ is the independence number of the strong product of n copies of G. We introduce an analogue of the Shannon capacity, namely $\kappa_s(G)=\inf_n \sqrt[n]{\kappa(G^n)}$, where $\kappa(G^n)$ is the independent domination number of the strong product of n copies of G. The Shannon capacity measures how rich a language can be, where the language is to be transmitted through a noisy channel. The parameter κ_s, on the other hand, measures how sparse such a language can be, if it is maximal with respect to inclusion.

AMS(MOS) subject classifications. 05C70, 90C05

Introduction. In 1956 Shannon [11] posed a problem in information theory which has led to many interesting graph theoretical questions. Suppose that we wish to transmit messages using letters from an alphabet $\mathscr{A}$. Due to noise in the channel, certain letters can be confused when transmitted. To be precise, when we transmit a letter, say j, any letter in a nonempty set S_j might be received. We say that two distinct letters, j and k, can be confused if $S_j \cap S_k$ is nonempty. Let b_n be the maximum number of n-letter words between which there is no confusion, where two distinct words can be confused iff their ith letters are the same or can be confused for each i. Shannon's problem is to evaluate $\sup_n \sqrt[n]{b_n}$.

Shannon's problem can be stated in graph theoretical terms. Consider the graph $G_{\mathscr{A}}=(V, E)$ which has one vertex for each letter in $\mathscr{A}$ and in which two vertices are adjacent iff the corresponding letters can be confused. Recall that the *independence number* of any graph H, which we will denote by $\alpha(H)$, is the maximum number of pairwise nonadjacent vertices of H. Thus $b_1=\alpha(G_{\mathscr{A}})$. Also, the *strong product* of the graphs $H_1=(V_1, E_1), \cdots, H_n=(V_n, E_n)$ is the graph $H_1\times H_2\times\cdots\times H_n$ on the vertex set $V_1\times V_2\times\cdots\times V_n$ in which two distinct vertices are adjacent iff their ith coordinates are equal or adjacent in H_i, for each i. It follows that $b_n=\alpha(G^n_{\mathscr{A}})$, for each n, where $G^n_{\mathscr{A}}$ is the strong product of n copies of $G_{\mathscr{A}}$, and that Shannon's problem is to evaluate

$$\alpha_s(G_{\mathscr{A}})\equiv\sup_n \sqrt[n]{\alpha(G^n_{\mathscr{A}})}.$$

The parameter α_s is known as the *Shannon capacity.*

The determination of the Shannon capacity of an arbitrary graph appears to be quite difficult. Indeed, the Shannon capacity of the 5-cycle, C_5, was not determined until 1978 [7], and the Shannon capacity of each odd cycle of length greater than 5 is still unknown. There has been a recent surge of interest in the study of the Shannon capacity, and numerous analogues have been introduced (see, for example, [5]). The purpose of this paper is to study another interesting analogue of the Shannon capacity.

A set S of vertices of the graph $G=(V, E)$ is *dominating* if every vertex in $V\backslash S$ is adjacent to some vertex in S. The *domination number* of G, denoted $\gamma(G)$, is the cardinality of a smallest dominating set in G, and the *independent domination number* of G, denoted $\kappa(G)$, is the smallest possible cardinality of a set which is both

* Received by the editors November 29, 1983, and in revised form August 13, 1984. The material in this paper is taken from the author's Ph.D. thesis which was written under the supervision of Dr. Pavol Hell.

† Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

independent and dominating. Notice that $\kappa(G)$ is the cardinality of a minimum cardinality maximal independent set in G. We define

$$\gamma_s(G) \equiv \inf_n \sqrt[n]{\gamma(G^n)} \quad \text{and} \quad \kappa_s(G) \equiv \inf_n \sqrt[n]{\kappa(G^n)}.$$

The parameter γ_s has been studied previously [8] and is, in fact, easy to evaluate (see below). We are interested in studying the parameter κ_s, which we refer to as the κ-capacity. Other work on the independent domination number of strong products of graphs can be found in [9]. It is straightforward to verify that κ is submultiplicative, and hence

$$\kappa_s(G) = \lim_{n\to\infty} \sqrt[n]{\kappa(G^n)}.$$

It is worth noting that, while the Shannon capacity of $G_{\mathscr{A}}$ yields an upper bound on the cardinality of a set of n-letter words which are pairwise nonconfusable, the κ-capacity of $G_{\mathscr{A}}$ yields a lower bound on the size of a maximal such set. In other words, the κ-capacity of $G_{\mathscr{A}}$ yields a lower bound on the size of a worst-possible set of pairwise nonconfusable n-letter words, where we are not obviously wasteful, i.e., where we would add another word to the set if possible.

Several related parameters are useful for the study of the Shannon capacity and its analogues γ_s and κ_s. For a given graph $G=(V, E)$ let $\mathscr{C}$ be the set of all cliques (i.e., maximal complete subgraphs) of G; and, for each vertex v, let $N[v]$ be the closed neighborhood of v, i.e., the set consisting of v together with all vertices adjacent to v. (For graph theoretical definitions which are not given here, see [1].)

Consider the following "fractional" versions of the parameters α, γ, and κ:

$$\alpha_f(G) \equiv \max\left\{\sum_{v\in V} x_v \colon \sum_{v\in C} x_v \leqq 1\ \forall C\in\mathscr{C}, x_v \geqq 0\ \forall v\in V\right\},$$

$$\gamma_f(G) \equiv \min\left\{\sum_{v\in V} x_v \colon \sum_{v\in N[u]} x_v \geqq 1\ \forall u\in V, x_v \geqq 0\ \forall v\in V\right\},$$

$$\kappa_f(G) \equiv \min\left\{\sum_{v\in V} x_v \colon \sum_{v\in C} x_v \leqq 1\ \forall C\in\mathscr{C}, \sum_{v\in N[u]} x_v \geqq 1\ \forall u\in V, \text{ and } x_v \geqq 0\ \forall v\in V\right\}.$$

We note that $\alpha_f(G)$ is usually referred to as the *Rosenfeld number* of G.

It is known that $\alpha_s(G) \leqq \alpha_f(G)$ [10] (cf. [11]) and $\gamma_s(G) = \gamma_f(G)$ [8] (cf. [5]), for every graph G. (Since linear programs can be solved in polynomial time [6], it follows that $\gamma_s(G)$ can be evaluated efficiently. Also, the quantity $\alpha(G)$ has been shown to be computable in polynomial time when G is perfect [4]. Since $\alpha(G) = \alpha_f(G)$ when G is perfect it follows that $\alpha_s(G)$ can be computed efficiently for a perfect graph G.) In light of these facts, one might expect that $\kappa_s(G) \geqq \kappa_f(G)$ for every graph G. Unfortunately, this fails to be true even for trees. The problem is that, unlike α_f and γ_f, κ_f is not multiplicative. (The multiplicativity of α_f and γ_f follow easily from the duality theorem of linear programming.) It is known that $\kappa(G) = \kappa_f(G)$ whenever G is chordal [3], and hence, whenever G is a tree. On the other hand, there are infinitely many trees T for which $\kappa_s(T) < \kappa(T)$ (see Theorem 1).

In the remainder of this paper we will present lower bounds on the κ-capacity and use them to evaluate the κ-capacity for several classes of graphs. As is common in the study of the Shannon capacity, we will use linear programming duality. However, we do so in a novel way (see the proof of Theorem 1).

The bounds. Notice that $\gamma(G) \leqq \kappa(G)$, for every graph G, and hence $\gamma_f(G) = \gamma_s(G) \leqq \kappa_s(G) \leqq \kappa(G)$. Since $\gamma_f(G)$ is easy to evaluate, it provides a practical lower bound on $\kappa_s(G)$. In general, this bound is not tight. However, it suffices to obtain exact values for several classes of graphs. For example, it is straightforward to verify that $\gamma_f(C_n) = n/3$ and $\kappa(C_n) = \lceil n/3 \rceil$, for each n. Thus, $\kappa_s(C_{3n}) = n$, for every n. As another example, we show that $\gamma_f(T) = \kappa_s(T) = \kappa(T)$ for each tree T which contains no edge each of whose ends has degree at least 3. (On the other hand, $\gamma_f(T) < \kappa_s(T) < \kappa(T)$ for the smallest tree T which does not satisfy these conditions—see Theorem 1.) It is known that $\gamma(T') = \gamma_f(T')$ for every tree T' [2]. Thus, it suffices to show that $\kappa(T) = \gamma(T)$. Choose a minimum cardinality dominating set D in T which minimizes the number of edges in the subgraph induced by D. If there are no edges in this subgraph, then $\kappa(T) = \gamma(T)$. Otherwise, there is an edge uv with $u, v \in D$. At least one of its ends, say u, has degree at most 2. Since $D \backslash \{u\}$ is not a dominating set, u has exactly one neighbor, say x, which is not dominated by $D \backslash \{u\}$. Thus $(D \cup \{x\}) \backslash \{u\}$ is a dominating set which induces one less edge than D induces, contradicting the choice of D.

We now present another lower bound on $\kappa_s(G)$ and use it to obtain exact values for the κ-capacity of certain trees which do not satisfy the condition $\gamma_f(G) = \kappa(G)$. This bound is, in general, tighter than $\gamma_f(G)$.

We define

$$\kappa_{fs}(G) \equiv \inf_n \sqrt[n]{\kappa_f(G^n)}.$$

By definition, $\kappa_{fs}(G) \leqq \kappa_s(G)$. Also, $\gamma_f(G) \leqq \kappa_{fs}(G)$, since γ_f is multiplicative.

For each $m \geqq 1$, let H_m be the tree on the vertices $u_0, u_1, \cdots, u_m, v_0, v_1, \cdots, v_m$ depicted in Fig. 1. Notice that for each m, $\gamma_f(H_m) = \gamma(H_m) = 2$ and $\kappa(H_m) = m + 1$.

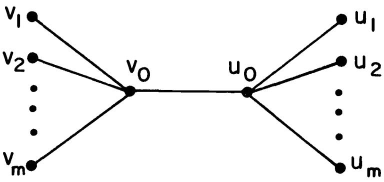


FIG. 1

THEOREM 1. *For each m, $\kappa_s(H_m) = 2\sqrt[n]{m}$, where H_m is the tree depicted in Fig.* 1.

Proof. Let $m \geqq 1$ and let

$$I_m = \{(u_0, u_i) \colon i = 1, 2, \cdots, m\} \cup \{(v_i, u_0) \colon i = 1, 2, \cdots, m\}$$
$$\cup \{(v_0, v_i) \colon i = 1, 2, \cdots, m\} \cup \{(u_i, v_0) \colon i = 1, 2, \cdots, m\}.$$

It is straightforward to verify that I_m is an independent dominating set in H_m^2 of cardinality $4m$. Consequently, $\kappa_s(H_m) \leqq 2\sqrt{m}$. It remains to show that $\kappa_s(H_m) \geqq 2\sqrt{m}$. It suffices to show that $\kappa_{fs}(H_m) \geqq 2\sqrt{m}$, i.e., that $\kappa_f(H_m^n) \geqq (4m)^{n/2}$ for every n. To do this, we will utilize the dual of the linear program associated with $\kappa_f(G)$, namely:

$$P(G)\colon \text{ Maximize } \sum_{u \in V} y_u - \sum_{C \in \mathscr{C}} z_C$$

$$\text{subject to: } \sum_{u \in N[v]} y_u - \sum_{C \ni v} z_C \leqq 1, \quad \forall v \in V,$$

$$y_u \geqq 0, \quad \forall u \in V,$$

$$z_C \geqq 0, \quad \forall C \in \mathscr{C}.$$

By the duality theorem of linear programming, it suffices to show that the optimal value of $P(H_m^n)$ is at least $(4m)^{n/2}$, for each n. We will do this by giving a canonical feasible solution to $P(H_m^n)$ of value $(4m)^{n/2}$, for every n.

Let C^* be the clique (edge) of H_m induced by $\{u_0, v_0\}$, and let $n > 0$. We will call a vertex of H_m^n a *p-vertex* if exactly p of its coordinates are endvertices of H_m, and we will call a clique of H_m^n a *p-clique* if it is a product of $n-p$ copies of C^* and p other cliques of H_m. (Notice that every clique of H_m^n is a product of n cliques of H_m, and vice versa.)

The following facts about H_m^n are easy to verify:

(1) There are exactly $(2m)^n$ n-vertices.

(2) There are exactly $n(2m)^{n-1}$ $(n-1)$-cliques.

(3) Each p-vertex is equal or adjacent to exactly m^{n-p} n-vertices.

(4) Each p-vertex is in exactly $(n-p)m^{n-p-1}$ $(n-1)$-cliques.

Using these facts we will show that the following is a feasible solution to $P(H_m^n)$ of value $(4m)^{n/2}$:

$$y_v = \begin{cases} m^{-(n/2)}(1+(n/2)\ln m) & \text{if } v \text{ is an } n\text{-vertex,} \\ 0 & \text{otherwise} \end{cases}$$

$$z_C = \begin{cases} m^{1-(n/2)}\ln m & \text{if } C \text{ is an } (n-1)\text{-clique,} \\ 0 & \text{otherwise.} \end{cases}$$

(Here, $\ln m$ denotes the natural logarithm of m.)

If v is a p-vertex then, by (3) and (4), we have

$$\begin{aligned} \sum_{u\in N[v]} y_u - \sum_{C\ni v} z_C &= m^{n-p}m^{-n/2}\left(1+\frac{n}{2}\ln m\right) - (n-p)m^{n-p-1}m^{1-(n/2)}\ln m \\ &= m^{(n/2)-p}\left(1+\frac{n}{2}\ln m\right) - m^{(n/2)-p}(n-p)\ln m \\ &= m^{(n/2)-p}\left(1-\left(\frac{n}{2}-p\right)\ln m\right) \\ &\leqq 1. \end{aligned}$$

Note that the last inequality follows from applying the first derivative test to the function f defined by $f(x) = m^x(1-x\ln m)$. Since each y_v and z_C is nonnegative it follows that this is a feasible solution to $P(H_m^n)$. Finally, using (1) and (2), we see that this solution yields an objective value of

$$\begin{aligned} \sum_{u\in V} y_u - \sum_{C\in \mathscr{C}} z_C &= (2m)^n m^{-n/2}\left(1+\frac{n}{2}\ln m\right) - n(2m)^{n-1}m^{1-(n/2)}\ln m \\ &= (4m)^{n/2}\left(1+\frac{n}{2}\ln m\right) - (4m)^{n/2}\frac{n}{2}\ln m \\ &= (4m)^{n/2}. \end{aligned}$$

□

Now, for each $m \geqq 1$ and $k \geqq 1$, let $G_{m,k}$ be the tree on the vertices $\{r, v_1, v_2, \cdots, v_k, w_1, w_2, \cdots, w_m, x_1, x_2, \cdots, x_{m^2}\}$ depicted in Fig. 2. Notice that $\gamma_f(G_{m,k}) = \gamma(G_{m,k}) = m+1$ and $\kappa(G_{m,k}) = m+k$ for all m and k.

THEOREM 2. *For each $m \geqq k \geqq 2$, $\kappa_s(G_{m,k}) = m+k$, where $G_{m,k}$ is the tree depicted in Fig. 2.*

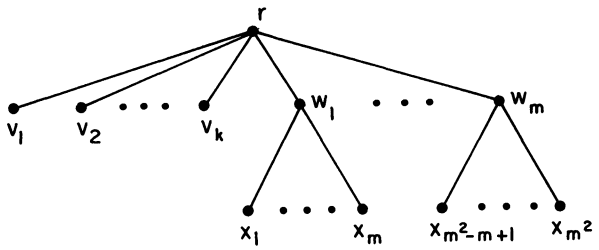


FIG. 2

Outline of proof. The proof is similar to that of Theorem 1. Since $\kappa(G_{m,k}) = m+k$, it suffices to show that $\kappa_s(G_{m,k}) \geqq m+k$. We establish this inequality by giving a canonical feasible solution to $P(G^n_{m,k})$ of value $(m+k)^n$, for each n.

We will call a vertex of $G^n_{m,k}$ an (a, b, c, d)-vertex if a of its coordinates are in $\{x_1, x_2, \cdots, x_{m^2}\}$, b of its coordinates are in $\{v_1, v_2, \cdots, v_k\}$, c of its coordinates are in $\{w_1, w_2, \cdots, w_m\}$, and d of its coordinates are r. Also, we will call a clique of $G^n_{m,k}$ an (a, b, c)-clique if it is a product of a cliques induced by some w_i and x_j, b cliques induced by r and some v_i, and c cliques induced by r and some w_i.

The following facts about $G^n_{m,k}$ are easy to verify:

(1) For each $a = 0, 1, 2, \cdots, n$, there are exactly $\binom{n}{a} m^{2a} k^{n-a}$ $(a, n-a, 0, 0)$-vertices.

(2) For each $a = 0, 1, 2, \cdots, n-1$ there are exactly $n\binom{n-1}{a} m^{2a+1} k^{n-a-1}$ $(a, n-a-1, 1)$-cliques.

(3) The number of $(a', n-a', 0, 0)$-vertices which are equal or adjacent to a specific (a, b, c, d)-vertex is

$$\begin{array}{ll} m^c k^d & \text{if } a' = a+c, \\ 0 & \text{otherwise.} \end{array}$$

(4) The number of $(a', n-a'-1, 1)$-cliques containing a specific (a, b, c, d)-vertex is

$$\begin{array}{ll} cm^{c-1}k^d & \text{if } a'+1 = a+c, \\ dm^{c+1}k^{d-1} & \text{if } a' = a+c, \text{ and} \\ 0 & \text{otherwise.} \end{array}$$

Using these facts, together with the fact that $k \leqq m$, one can show that the following is a feasible solution to $P(G^n_{m,k})$ of value $(m+k)^n$:

$$y_v = \begin{cases} m^{-a}(1 + a \ln m) & \text{if } v \text{ is an } (a, n-a, 0, 0)\text{-vertex, } a = 0, 1, 2, \cdots, n, \\ 0 & \text{otherwise,} \end{cases}$$

$$z_C = \begin{cases} m^{-a} \ln m & \text{if } C \text{ is an } (a, n-a-1, 1)\text{-clique, } a = 0, 1, 2, \cdots, n-1, \\ 0 & \text{otherwise.} \end{cases}$$

We leave the details to the interested reader. □

Concluding remarks. Notice that every graph G considered in the preceding section satisfies

$$\kappa_{fs}(G) = \kappa_s(G). \tag{3.1}$$

If (3.1) were true in general then the κ-capacity of many other graphs could be evaluated. For example, this would imply that $\kappa_s(C_n) = n/3$ for every n. Although we know of no counter-example to (3.1), we suspect that it does not hold in general. Indeed, we conjecture that $\kappa_s(C_4) = \sqrt[3]{4}$. To see that $4/3 \leqq \kappa_s(C_4) \leqq \sqrt[3]{4}$, notice that $\gamma_f(C_4) = 4/3$. On the other hand, if we number the vertices of C_4 by $0, 1, 2, 3$ so that $[i, j]$ is an edge iff $i+j$ is odd, then $\{(0, 0, 0), (1, 2, 3), (2, 3, 1), (3, 1, 2)\}$ is an independent dominating set in C_4^3.

It would be interesting to determine classes of graphs which satisfy (3.1). We conjecture that (3.1) holds for trees.

Since γ_f appears to be a poor lower bound on κ_s, and κ_{fs} is not multiplicative, the most obvious problem which comes to light is to find a good multiplicative lower bound on κ.

Acknowledgment. We would like to thank Dr. Pavol Hell who suggested the problem of studying the κ-capacity, and who supervised the Ph.D. thesis from which this paper was extracted.

REFERENCES

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

[2] M. FARBER, *Domination and duality in weighted trees*, Congressus Numerantium, 33 (1981), pp. 3-13.

[3] ———, *Independent domination in chordal graphs*, Oper. Res. Letters, 1 (1982), pp. 134-138.

[4] M. GRÖTSCHEL, L. LOVÁSZ AND A. SCHRIJVER, *The ellipsoidal method and its consequences in combinatorial optimization*, Combinatorica, 1 (1981), pp. 169-197.

[5] P. HELL AND F. S. ROBERTS, *Analogues of the Shannon capacity of a graph*, Ann. Discrete Math., 12 (1982), pp. 155-168.

[6] L. G. KHACHIAN, *A polynomial algorithm in linear programming*, Soviet Math. Dokl., 20 (1979), pp. 191-194.

[7] L. LOVÁSZ, *On the Shannon capacity of a graph*, IEEE Trans. Inform. Theory, IT-25 (1979), pp. 1-7.

[8] R. J. MCELIECE AND E. C. POSNER, *Hide and seek, data storage, and entropy*, Ann. Math. Stat., 42 (1971), pp. 1706-1716.

[9] J. D. MCFALL AND R. J. NOWAKOWSKI, *On strong independence and the strong product of graphs*, in preparation.

[10] M. ROSENFELD, *On a problem of C. E. Shannon in graph theory*, Proc. Amer. Math. Soc., 18 (1967), pp. 315-319.

[11] C. E. SHANNON, *The zero-error capacity of a noisy channel*, IRE Trans. Inform. Theory, IT-2 (1956), pp. 8-19.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

009

PROBABILITIES FOR INTERSECTING SYSTEMS AND RANDOM SUBSETS OF FINITE SETS*

P. C. FISHBURN†, P. FRANKL‡, D. FREED§, J. C. LAGARIAS† AND A. M. ODLYZKO†

Abstract. Let $\mathcal{F}_k$ be a family of subsets of $\{1, 2, \cdots, n\}$, each two of which have at least k elements in common, and let S be a random subset (sample) of $\{1, 2, \cdots, n\}$ obtained by choosing each $i \leqq n$ independently with probability p_i. Assuming that $1 > p_1 \geqq p_2 \geqq \cdots \geqq p_n > 0$, we investigate the problem of determining an $\mathcal{F}_k$ that maximizes the probability that at least one of the sets in $\mathcal{F}_k$ will be included in S.

Complete solutions are obtained for the following cases: for $k=1$, $p_i = p$ for all i, or variable p_i with $\frac{1}{2} \geqq p_2$ or $p_n > \frac{1}{2}$; for $k \geqq 2$, $p_i = p \geqq \frac{1}{2}$ for all i, and small p_{k+1}. A partial solution is given for $k=2$ when $p_i = p$ for all i.

1. Introduction. Let $\mathcal{F}_k$ be a family of subsets of $\mathbf{n} = \{1, 2, \cdots, n\}$ for which $1 \leqq k < n$ and

$$(1) \qquad |A \cap B| \geqq k \quad \text{for all } A, B \in \mathcal{F}_k,$$

and let $\mathbf{p} = (p_1, p_2, \cdots, p_n)$ be a probability vector with

$$1 > p_1 \geqq p_2 \geqq \cdots \geqq p_n > 0.$$

Our aim is to determine intersecting families $\mathcal{F}_k$ that maximize the probability $P(\mathcal{F}_k, \mathbf{p})$ that at least one member of $\mathcal{F}_k$ will be included in a random subset S of $\mathbf{n}$ that independently contains each $i \leqq n$ with probability p_i. We thus add a probabilistic dimension to the theory of intersecting systems initiated by Erdös, Ko and Rado (EKR) [3] and surveyed recently by Deza and Frankl [2].

Let

$$\mathcal{F}^+ = \{B \subseteq \mathbf{n} \colon A \subseteq B \text{ for some } A \in \mathcal{F}\}.$$

Clearly, $\mathcal{F}_k^+$ satisfies (1) and $P(\mathcal{F}_k^+, \mathbf{p}) = P(\mathcal{F}_k, \mathbf{p})$. It is also easily shown that if $\mathbf{p} \geqq \mathbf{p}'$ ($p_i \geqq p_i'$ for all i), then $P(\mathcal{F}_k, \mathbf{p}) \geqq P(\mathcal{F}_k, \mathbf{p}')$. Unless it is stated otherwise, all $\mathcal{F}_k$ in what follows will be maximal intersecting systems, i.e., $\mathcal{F}_k = \mathcal{F}_k^+$ and no set can be added to $\mathcal{F}_k$ without violating (1).

The simplest case of probabilistic EKR theory has $k=1$ and $\mathbf{p}$ constant, say $p_i = p$ for all i. To illustrate, suppose each card in a 52-card deck is to be independently chosen with probability p for a sample S of the deck. $\mathcal{F}_1$ is a maximal family of subdecks, each two of which have at least one card in common, and $P(\mathcal{F}_1, \mathbf{p})$ is the probability that some subdeck in $\mathcal{F}_1$ will be included in S. The following results were first established indirectly by Ahlswede and Katona [1] in their Theorem 4.2 and were independently derived by the present authors. If $p < \frac{1}{2}$, then the maximum of $P(\mathcal{F}_1, \mathbf{p})$ over $\mathcal{F}_1$ equals p, and this value is realized when $\mathcal{F}_1$ consists of a one-card subdeck and all of its supersets. If $p > \frac{1}{2}$, then max $P(\mathcal{F}_1, \mathbf{p}) > p$, and any maximizing $\mathcal{F}_1$ contains all subdecks with more than half the cards plus half the subdecks with exactly 26 cards. If $p = \frac{1}{2}$, then max $P(\mathcal{F}_1, \mathbf{p}) = \frac{1}{2}$, and this is realized by every maximal $\mathcal{F}_1$.

The results for constant $\mathbf{p}$ and $k=1$ are discussed along with two variable-$\mathbf{p}$ generalizations for $k=1$ in the next section. The first generalization has $\frac{1}{2} \geqq p_2$; the second has $p_n > \frac{1}{2}$. Definitive results for $k=1$ are not presently known for other cases.

* Received by the editors July 26, 1983.

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

‡ Centre National de la Recherche Scientifique, Paris, and Consultant, AT & T Bell Laboratories.

§ University of California, Berkeley, California 94720. This work was done while this author was a summer employee at AT & T Bell Laboratories, Murray Hill, New Jersey 07974.

Later sections consider $k \geqq 2$. Section 3 shows that a P-maximizing $\mathcal{F}_k$ consists of $\{1, 2, \cdots, k\}$ and its supersets if p_{k+1} through p_n are small. Section 4 proves that collections of large subsets of $\mathbf{n}$ maximize P when $\mathbf{p}$ is constant and $p \geqq \frac{1}{2}$. Section 5 gives a partial result for all constant $\mathbf{p}$ when $k = 2$. It is hoped that further research will add substantially to these results.

Several notations apply throughout the paper. If not indicated otherwise, A and B are subsets of $\mathbf{n}$, and $A^c = \mathbf{n} \backslash A$. $\mathcal{F}_k^*$ denotes an $\mathcal{F}_k$ that maximizes $P(\mathcal{F}_k, \mathbf{p})$, and $\mathcal{F}_k^{**}$ is an $\mathcal{F}_k$ that uniquely maximizes $P(\mathcal{F}_k, \mathbf{p})$.

2. Probabilistic EKR theory for $k = 1$. We assume $k = 1$ throughout this section and will omit the subscript on $\mathcal{F}_1$ for convenience. Given $k = 1$, a maximal $\mathcal{F}$ must contain A or A^c, else there would be $B, C \in \mathcal{F}$ with $B \subseteq A^c$ and $C \subseteq A$, thus contradicting $B \cap C \neq \emptyset$. Since both A and A^c cannot be in $\mathcal{F}$, we have the well-known

LEMMA 1. *Every maximal $\mathcal{F}$ contains exactly 2^{n-1} subsets of n.*

We begin with the constant-$\mathbf{p}$ cases. Since all S's are equally likely when $p_i = \frac{1}{2}$ for all i, every maximal $\mathcal{F}$ is an $\mathcal{F}^*$ for this case, and $P(\mathcal{F}^*, (\frac{1}{2}, \cdots, \frac{1}{2})) = 2^{n-1}/2^n = \frac{1}{2}$. Other values of p are covered by

THEOREM 1. *Suppose $p_i = p$ for all i. Then*

$$p < \tfrac{1}{2} \Rightarrow \mathcal{F}^* = \{\{i\}\}^+;$$

$$[p > \tfrac{1}{2}, n \text{ odd}] \Rightarrow \mathcal{F}^{**} = \left\{A \colon |A| \geqq \frac{n+1}{2}\right\};$$

$$[p > \tfrac{1}{2}, n \text{ even}]$$

$$\Rightarrow \mathcal{F}^* = \left\{A \colon |A| > \frac{n}{2}\right\} \cup \left\{\text{half of the } A \text{ with } |A| = \frac{n}{2}, \text{ one from each } \{A, A^c\} \text{ pair}\right\}.$$

The values of $P(\mathcal{F}^*, \mathbf{p})$ are easily determined from the theorem and reveal an interesting discontinuity in the limit:

$$\lim_{n \to \infty} P(\mathcal{F}^*, \mathbf{p}) = \begin{cases} p & \text{if } p \leqq \frac{1}{2}, \\ 1 & \text{if } p > \frac{1}{2}. \end{cases}$$

When $p > \frac{1}{2}$, the probability that S contains more than half the elements in $\mathbf{n}$ approaches 1 as $n \to \infty$.

Although Ahlswede and Katona's results [1] yield a proof of Theorem 1, we present a full proof to show the application of a basic result of Erdös, Ko and Rado [3]. Ingenious short proofs of the following are given by Katona [8] and Greene and Kleitman [6].

LEMMA 2 (Erdös-Ko-Rado). *For each $1 \leqq t < n/2$, the largest collection of t-element subsets of $\mathbf{n}$ that are pairwise nondisjoint has cardinality $\binom{n-1}{t-1}$. This maximum is realized by the collection of all t-element subsets of $\mathbf{n}$ that contain a fixed $i \leqq n$, and this is unique up to the choice of i.*

To prove Theorem 1, let a_t be the number of t-sets (t-element subsets) in $\mathcal{F}$. By the proof of Lemma 1,

$$a_t + a_{n-t} = \binom{n}{t}, \qquad t = 1, \cdots, n,$$

with $\sum a_t = 2^{n-1}$. Let $w_t = p^t(1-p)^{n-t}$ so that

$$P(\mathcal{F}, \mathbf{p}) = \sum a_t w_t \quad \text{with } \mathbf{p} = (p, \cdots, p).$$

If $p<\frac{1}{2}$, then $w_1>w_2>\cdots>w_n$, so, by Lemma 2 and $a_t+a_{n-t}=\binom{n}{t}$, $a_tw_t+a_{n-t}w_{n-t}$ for $t<n/2$ is maximized if and only if the t-sets in $\mathscr{F}$ are those that contain a fixed i. It then follows that $\sum a_tw_t$ is maximized if and only if some i is in every $A\in\mathscr{F}$.

If $p>\frac{1}{2}$, then $w_n>w_{n-1}>\cdots>w_1$, and therefore $\sum a_tw_t$ is maximized by making the a_t for $t>n/2$ as large as possible, namely $\binom{n}{t}$. If n is odd, this yields the unique maximizer $\mathscr{F}^{**}$ shown in the theorem. If n is even, then all maximal $\mathscr{F}$ have $\binom{n}{n/2}/2$ $(n/2)$-sets, and any such $\mathscr{F}$ that contains every A with $|A|>n/2$ is a maximizer of P. This completes the proof of Theorem 1.

We now consider variable $\mathbf{p}$ with $1>p_1\geqq p_2\geqq\cdots\geqq p_n>0$, as we assume throughout. Our first result in this case pertains to relatively small p_i.

THEOREM 2. *Suppose* $\frac{1}{2}\geqq p_2$. *Then* $\mathscr{F}^*=\{\{1\}\}^+$, *and* $\mathscr{F}^{**}=\{\{1\}\}^+$ *iff* $p_1>p_2$.

Proof. Given $\frac{1}{2}\geqq p_2$, let $\mathscr{A}=\{\{1\}\}^+$ and let $\mathscr{F}$ be any other maximal intersecting family. Consider $\mathbf{p}'=(p_2,p_2,\cdots,p_2)$. By Theorem 1, $\mathscr{A}$ is an $\mathscr{F}^*$ for $\mathbf{p}'$, so $P(\mathscr{A},\mathbf{p}')\geqq P(\mathscr{F},\mathbf{p}')$. Let $\mathbf{p}''=(p_2,p_2,p_3,\cdots,p_n)$. Since $P(\mathscr{A})$ is not affected by values of the p_i beyond the initial value, but $P(\mathscr{F})$ cannot increase as those values decrease, $P(\mathscr{A},\mathbf{p}'')=P(\mathscr{A},\mathbf{p}')\geqq P(\mathscr{F},\mathbf{p}')\geqq P(\mathscr{F},\mathbf{p}'')$, so $P(\mathscr{A},\mathbf{p}'')\geqq P(\mathscr{F},\mathbf{p}'')$. If $p_1=p_2$, then $\mathbf{p}=\mathbf{p}''$, so $\mathscr{A}$ and—by symmetry—$\{\{2\}\}^+$ are $\mathscr{F}^*$'s for $\mathbf{p}$.

Suppose then that $p_1>p_2$. As we change from $\mathbf{p}''$ to $\mathbf{p}$, $P(\mathscr{A})$ increases from p_2 to p_1. If $B\in\mathscr{F}$, then $Pr(B\subseteq S)$ does not change when $\mathbf{p}''$ changes to $\mathbf{p}$ if $1\notin B$, and it increases by a factor of p_1/p_2 if $1\in B$. Since $\{1\}^c\in\mathscr{F}$, it follows that

$$P(\mathscr{A},\mathbf{p})=p_1=p_2(p_1/p_2)>P(\mathscr{F},\mathbf{p}).$$

Thus $\mathscr{A}$ is the unique maximizer if $p_1>p_2$. □

Our second generalization of Theorem 1 for $k=1$ takes all $p_i>\frac{1}{2}$. In this case, each $\mathscr{F}^*$ is determined by the greedy algorithm that chooses A from $\{A,A^c\}$ if $P(A)>P(A^c)$ and chooses either if $P(A)=P(A^c)$, where

$$P(A)=Pr(A\subseteq S)\text{ when }\mathbf{p}\text{ applies.}$$

THEOREM 3. *Suppose* $p_n>\frac{1}{2}$. *Then* $\mathscr{F}^*=\{A\colon P(A)>P(A^c)\}\cup\{$*half of the* A *with* $P(A)=P(A^c)$, *one from each such* $\{A,A^c\}$ *pair*$\}$.

Remark. As in the final part of Theorem 1, either A or A^c can be chosen for $\mathscr{F}^*$ in Theorem 3 when $P(A)=P(A^c)$. The ensuing proof shows that intersections of chosen sets will be nonempty. If $p_n>\frac{1}{2}$ is replaced by $p_n\geqq\frac{1}{2}$, slight modifications in the proof show that $\mathscr{F}^*$ equals $\{A\colon P(A)>P(A^c)\}$ plus one set from each $\{A,A^c\}$ that has $P(A)=P(A^c)$. The latter choices may require explicit consideration of nonempty intersections if some $p_i=\frac{1}{2}$.

Proof of Theorem 3. Given $p_n>\frac{1}{2}$, let $\mathscr{F}$ be an $\mathscr{F}^*$ as designated in the theorem. Contrary to the theorem's conclusion, suppose A, $B\in\mathscr{F}$ and $A\cap B=\varnothing$. Let $C=\mathbf{n}\backslash(A\cup B)$. This is not empty, since otherwise $B=A^c$, contrary to the definition of $\mathscr{F}$. By that definition, $P(A)\geqq P(A^c)$ and $P(B)\geqq P(B^c)$, so

$$\frac{P(A)P(B)}{P(A^c)P(B^c)}\geqq 1.$$

However, the P ratio here equals $\prod_{i\in C}[(1-p_i)/p_i]^2$, which is strictly less than 1 since $p_i>\frac{1}{2}$, so a contradiction obtains. Hence A, $B\in\mathscr{F}$ implies $A\cap B\neq\varnothing$. □

In contrast to the final conclusions of Theorem 1, where $\mathscr{F}^*$ consists entirely of A with $|A|\geqq n/2$ when p is constant and $p>\frac{1}{2}$, suppose $p_1=1-\varepsilon$ and $p_i=\frac{1}{2}+\varepsilon$ for all $i\geqq 2$, ε positive and small. Theorem 3 then implies that $\mathscr{F}^{**}=\{\{1\}\}^+$. However, if ε is held fixed and n increases, we will eventually get to an n where $\mathscr{F}^*$ no longer contains $\{1\}$.

3. Small p_{k+1} for $k \geqq 2$. We assume $k \geqq 2$ henceforth. The following theorem generalizes Theorem 2 for small p_{k+1}. The p_i for $i \leqq k$ can of course be large.

THEOREM 4. *Suppose $2 \leqq k < n$. If either*

$$2 \leqq k \leqq 14 \quad \textit{and} \quad p_{k+1} \leqq \frac{1}{2(k+1)},$$

or

$$k \geqq 15 \qquad \textit{and} \quad p_{k+1} \leqq \frac{1}{k+1},$$

then $P(\mathcal{F}_k^*, \mathbf{p}) \leqq p_1 p_2 \cdots p_k$.

We suspect the conclusion also holds for $2 \leqq k \leqq 14$ when $p_{k+1} \leqq 1/(k+1)$, but lack proof. At any rate, $P(\mathcal{F}_k^*, \mathbf{p}) \leqq p_1 p_2 \cdots p_k$ clearly implies that $\mathcal{F}_k^* = \{\{1, 2, \cdots, k\}\}^+$ and that this is $\mathcal{F}_k^{**}$ iff $p_k > p_{k+1}$. When $p_{k+1} > 1/(k+1)$, the conclusion of the theorem is not generally true. For example, if $p_1 = p_2 = \cdots = p_{k+2} = p > 1/(k+1)$ and $\mathcal{F}_k = \{A: |A \cap \{1, \cdots, k+2\}| \geqq k+1\}$ then

$$P(\mathcal{F}_k, \mathbf{p}) = (k+2)p^{k+1}(1-p) + p^{k+2} = p^{k+1}(k+2-p(k+1)) > p^k.$$

Our method of proving Theorem 4 is essentially the same as a method used by Frankl and Füredi [5]. The main tool is the following extension of the Erdös-Ko-Rado theorem.

LEMMA 3 (Frankl [4]). *Suppose $k \leqq r \leqq n$ and $\mathcal{F}_{k,r}$ is a maximum-cardinality set of r-sets for which $|A \cap B| \geqq k$ for all A, B in the set. Let*

$$\mathcal{F}_{k,r}^0 = \{A: |A| = r \textit{ and } \{1, 2, \cdots, k\} \subseteq A\},$$
$$\mathcal{F}_{k,r}^1 = \{A: |A| = r \textit{ and } |A \cap \{1, \cdots, k+2\}| \geqq k+1\}.$$

Then, up to permutations on $\{1, \cdots, n\}$,

$$[2 \leqq k \leqq 14,\ n \geqq c_k(r-k+1)(k+1) \textit{ for some constant } 1 < c_k < 2 \textit{ that depends only on } k] \Rightarrow \mathcal{F}_{k,r} = \mathcal{F}_{k,r}^0;$$

$$[k \geqq 15,\ n > (r-k+1)(k+1)] \Rightarrow \mathcal{F}_{k,r} = \mathcal{F}_{k,r}^0;$$

$$[k \geqq 15,\ n = (r-k+1)(k+1)] \Rightarrow \mathcal{F}_{k,r} \in \{\mathcal{F}_{k,r}^0, \mathcal{F}_{k,r}^1\};$$

$$[k \geqq 15,\ c_k(r-k+1)(k+1) \leqq n < (r-k+1)(k-1) \textit{ for some constant } c_k < 1 \textit{ that depends only on } k] \Rightarrow \mathcal{F}_{k,r} = \mathcal{F}_{k,r}^1.$$

We also use two other results.

FACT 1. *If Theorem 4 is true whenever $p_{k+1} = \cdots = p_n$, then it is true in general.*

Proof. Let $\mathcal{F}_k^0 = \{\{1, \cdots, k\}\}^+$. Let $\mathbf{p} = (p_1, \cdots, p_{k+1}, p_{k+2}, \cdots, p_n)$ and $\mathbf{p}' = (p_1, \cdots, p_{k+1}, p_{k+1}, \cdots, p_{k+1})$. If Theorem 4 is true whenever $p_{k+1} = \cdots = p_n$, then, since $p_{k+1} \geqq p_{k+2} \geqq \cdots \geqq p_n$, and in view of the second paragraph in the introduction,

$$P(\mathcal{F}_k^0, \mathbf{p}) = p_1 \cdots p_k = P(\mathcal{F}_k^0, \mathbf{p}') \geqq P(\mathcal{F}_k, \mathbf{p}') \geqq P(\mathcal{F}_k, \mathbf{p}). \qquad \square$$

FACT 2. *Given $\mathbf{p} = (p_1, \cdots, p_k, p, \cdots, p)$ and $P_n^* = P(\mathcal{F}_k^*, \mathbf{p}$ with n components), it follows that $P_{n+1}^* \geqq P_n^*$ and hence that $\lim_{n\to\infty} P_n^*$ exists.*

Proof. Just consider the trivial extension of $\mathcal{F}_k^*$ at n to $n+1$ by adjoining to $\mathcal{F}_k^*$ the set $\{A \cup \{n+1\}: A \in \mathcal{F}_k^*\}$. □

By Fact 1, it suffices to show that Theorem 4 holds when $p_{k+1} = \cdots = p_n = p$, so assume henceforth that $\mathbf{p} = (p_1, \cdots, p_k, p, \cdots, p)$. Then Fact 2 yields the desired result of

$$\lim_{n\to\infty} P_n^* = p_1 p_2 \cdots p_k.$$

We prove this for $k \geqq 15$ in Theorem 4. The proof for $2 \leqq k \leqq 14$ is simpler since it involves only the central limit theorem.

Given $k \geqq 15$ and $p_{k+1} = p \leqq 1/(k+1)$, let $\mathscr{F}_k$ be an arbitrary maximal k-intersecting system, and let $\mathscr{F}_{k,s} = \{A \in \mathscr{F}_k \colon |A| = s\}$ for $s \geqq k$. Also let $a_s = |\mathscr{F}_{k,s}|$, $b = p_1 \cdots p_k$, and $f(k, n) = [k-1+n/(k+1)]$. Then, for large n, and in view of the first result of Lemma 3 for $k \geqq 15$, namely $\binom{n-k}{s-k} = |\mathscr{F}^0_{k,s}| \geqq a_s$ for $s < f(k, n)$, and $c_k < 1$ in the final result of the lemma, we have

$$\begin{aligned} P(\mathscr{F}_k, \mathbf{p}) &= \sum_{s=k}^{n} P(\mathscr{F}_{k,s}, \mathbf{p}) \leqq b \sum_{s=k}^{n} a_s p^{s-k}(1-p)^{n-s} \\ &= b - b \sum_{s=k}^{n} \left[\binom{n-k}{s-k} - a_s\right] p^{s-k}(1-p)^{n-s} \\ &\leqq b + bp^{-k} \sum_{s=f(k,n)}^{f(k,n)+[n^{2/3}]} \big||\mathscr{F}^1_{k,s}| - |\mathscr{F}^0_{k,s}|\big| p^s (1-p)^{n-s} \\ &\quad + bp^{-k} \sum_{s>f(k,n)+[n^{2/3}]} \binom{n}{s} p^s(1-p)^{n-s}. \end{aligned}$$

The final sum vanishes as $n \to \infty$ by the central limit theorem, and it is easily checked that

$$\begin{aligned} \sum_{s=f(k,n)}^{f(k,n)+[n^{2/3}]} (|\mathscr{F}^1_{k,s}| - |\mathscr{F}^0_{k,s}|) p^s(1-p)^{n-s} &= \sum \binom{n}{s} O(n^{-1/3}) p^s (1-p)^{n-s} \\ &= O(n^{-1/3}), \end{aligned}$$

which also vanishes in the limit. Therefore $\lim P(\mathscr{F}^*_k, \mathbf{p}) \leqq b$. □

4. Large p_i for $k \geqq 2$. Our main result for $k \geqq 2$ and $p_n \geqq \frac{1}{2}$ returns to the constant-$\mathbf{p}$ context of Theorem 1. We shall comment briefly on variable $\mathbf{p}$ shortly.

The following theorem of Katona [7] and Kleitman [9] illustrates another facet of standard EKR theory. Let

$$\mathscr{G}_k = \begin{cases} \left\{A \colon |A| \geqq \dfrac{n+k}{2}\right\} & \text{if } n+k \text{ is even,} \\[2ex] \left\{A \colon |A \cap \{1, \cdots, n-1\}| \geqq \dfrac{n-1+k}{2}\right\} & \text{if } n+k \text{ is odd.} \end{cases}$$

LEMMA 4 (Katona, Kleitman). *Suppose $\mathscr{F}_k$ has maximum cardinality. If $n+k$ is even, $\mathscr{F}_k = \mathscr{G}_k$. If $n+k$ is odd, then $\mathscr{F}_k$ is either $\mathscr{G}_k$ or an isomorph of $\mathscr{G}_k$ obtained by replacing $\{1, \cdots, n-1\}$ in its definition by any other $(n-1)$-set in* $\mathbf{n}$.

THEOREM 5. *Suppose $k \geqq 2$ and $p_i = p \geqq \frac{1}{2}$ for all i. Then $\mathscr{F}^{**}_k = \mathscr{G}_k$ if $n+k$ is even, and $\mathscr{F}^*_k = \mathscr{G}_k$ if $n+k$ is odd.*

Proof. Immediate from Lemma 4 and the fact that $p^t(1-p)^{n-t}$ is nondecreasing in t when $p \geqq \frac{1}{2}$. □

Two factors suggest that the variable-$\mathbf{p}$ case for $p_n \geqq \frac{1}{2}$ is more complex. First, specific applications of Theorem 3 for $k = 1$ show that $\mathscr{F}^*$ can vary considerably as the $p_i \geqq \frac{1}{2}$ change. Second, unlike $k = 1$, maximal $\mathscr{F}_k$ for $k \geqq 2$ can have different cardinalities. For example, with $\mathcal{A}_k = \{\{1,2,\cdots,k\}\}^+$,

$$\lim_{n\to\infty} |\mathcal{G}_k|/|\mathcal{A}_k| = 2^{k-1}.$$

However, $\mathcal{G}_k$'s size advantage over $\mathcal{A}_k$ may be offset under P if p_1 through p_k are near 1 while later p_i are near $\frac{1}{2}$.

We illustrate this for $k=2$ with $p_1 = p_2 = 1-\varepsilon$ and $p_3 = \cdots = p_n = \frac{1}{2} + \delta$. Regardless of n, $P(\mathcal{A}_2, \mathbf{p}) = (1-\varepsilon)^2$. If δ is fixed, then $P(\mathcal{G}_2, \mathbf{p})$ approaches 1 as n gets large, so that

$$P(\mathcal{G}_2, \mathbf{p}) > P(\mathcal{A}_2, \mathbf{p}) \quad \text{for } n > n_0(\varepsilon, \delta).$$

However, if δ varies with n and approaches 0 sufficiently rapidly, then $P(\mathcal{G}_2, \mathbf{p})$ will approach $\frac{1}{2}$ as n gets large, so that

$$P(\mathcal{A}_2, \mathbf{p}) > P(\mathcal{G}_2, \mathbf{p}) \quad \text{for large } n.$$

5. Constant p for $k=2$. We conclude with observations for constant $\mathbf{p}$ and $k=2$ that augment the small-p_i results of Theorem 4 and the large-p_i results of Theorem 5. Our prime question is what happens to $\mathcal{F}_2^*$ for p between ε_2 and $\frac{1}{2}$. The answer we give is incomplete since it deals only with a small number of maximal $\mathcal{F}_2$. However, it does suggest what the general $\mathcal{F}_2^*$ solution may look like when $k=2$ and $\mathbf{p}$ is constant.

For each $1 \leqq t \leqq [n/2]$ (the integer part of $n/2$), let

$$\mathcal{F}_2(t) = \{A: |A| = t+1 \text{ and } A \subseteq \{1,2,\cdots,2t\}\}^+,$$

the family of $(t+1)$-sets in $\mathbf{2t}$ and their supersets in $\mathbf{n}$. It is easily checked that each $\mathcal{F}_2(t)$ is a maximal $\mathcal{F}_2$ set. In previous notation, $\mathcal{F}_2(1)$ is $\mathcal{A}_2$, and $\mathcal{F}_2([n/2])$ is $\mathcal{G}_2$. Our partial answer to the question of the preceding paragraph is

THEOREM 6. *Suppose $p_i = p$ for all i, and $\mathcal{F}_2$ is restricted to $\{\mathcal{F}_2(1), \cdots, \mathcal{F}_2([n/2])\}$. Then $P(\mathcal{F}_2, p)$ is uniquely maximized by*

$$\mathcal{F}_2(t) \quad \textit{if} \quad \frac{t-1}{2t-1} < p < \frac{t}{2t+1} \textit{ for } t = 1,2,\cdots,[n/2];$$

$$\mathcal{F}_2([n/2]) \quad \textit{if} \quad \frac{[n/2]-1}{2[n/2]-1} < p < 1.$$

Thus, as p increases with $\mathcal{F}_2$ confined to the $\mathcal{F}_2(t)$, the optimal $\mathcal{F}_2(t)$ changes from $\mathcal{F}_2(1)$ to $\mathcal{F}_2(2)$, then to $\mathcal{F}_2(3)$, and so on, up to $\mathcal{F}_2([n/2])$ just before $p = \frac{1}{2}$. The only p where we definitely know that the designated $\mathcal{F}_2(t)$ is an $\mathcal{F}_2^*$ are $p \leqq \varepsilon_2$ and $p \geqq \frac{1}{2}$.

Our proof of Theorem 6 is based on a lemma that seems interesting in its own right. For the lemma let

$$\mathcal{H}(t) = \{A: |A| = t+1 \text{ and } A \subseteq \{1,2,\cdots,2t\}\}$$

for $t = 1,2,\cdots$ without specific reference to n, and let $P_t(p)$ denote the probability that a random S chosen from $\{1,2,\cdots,2t\}$ with probability p for each $i \leqq 2t$ will include a set in $\mathcal{H}(t)$.

LEMMA 5. *For each $t \geqq 1$,*

$$P_{t+1}(p) - P_t(p) = \binom{2t}{t+1} p^{t+1}(1-p)^t[(2t+1)p - t]/t.$$

Proof (outline). The lemma claims that

$$\sum_{k=t+2}^{2t+2}\binom{2t+2}{k}p^k(1-p)^{2t+2-k}-\sum_{k=t+1}^{2t}\binom{2t}{k}p^k(1-p)^{2t-k}$$
$$=\binom{2t}{t+1}p^{t+1}(1-p)^t[(2t+1)p-t]/t.$$

This can be verified by expanding both sides to obtain polynomials in p and showing that the coefficients of p^s are the same on both sides for $s=t+1,\cdots,2t+2$. The latter step makes extensive use of the identity

$$\sum_{j=0}^{a}(-1)^j\binom{b+1}{j}=(-1)^a\binom{b}{a}.$$

We omit the details. □

Proof of Theorem 6. Since $\mathcal{F}_2(t)=\mathcal{H}(t)^+$for $t\leqq[n/2]$, it follows that $P(\mathcal{F}_2(t),\mathbf{p})=P_t(p)$. The identity of Lemma 5 shows that the curves of $P_t(p)$ and $P_{t+1}(p)$ cross at

$$p=\frac{t}{2t+1},$$

with $P_t(p)>P_{t+1}(p)$ when $p<t/(2t+1)$, and $P_{t+1}(p)>P_t(p)$ when $p>t/(2t+1)$. The theorem follows directly from these observations. □

REFERENCES

[1] R. AHLSWEDE AND G. O. H. KATONA, *Contributions to the geometry of Hamming spaces*, Discrete Math., 17 (1977), pp. 1-22.

[2] M. DEZA AND P. FRANKL, *Erdös-Ko-Rado theorem—20 years later*, unpublished manuscript, 1981.

[3] P. ERDÖS, C. KO AND R. RADO, *Intersection theorems for systems of finite sets*, Quart. J. Math. Oxford (2), 12 (1961), pp. 313-320.

[4] P. FRANKL, *The Erdös-Ko-Rado theorem is true for n − ckt*, in Colloquia Math. Soc. Janos Bolyai 18, A. Hajnal and V. T. Sós, eds., Combinatorics, Vol. 1, North-Holland, Amsterdam, 1978, pp. 365-375.

[5] P. FRANKL AND Z. FÜREDI, *The Erdös-Ko-Rado theorem for integer sequences*, this Journal, 1 (1980), pp. 376-381.

[6] C. GREENE AND D. J. KLEITMAN, *Proof techniques in the theory of finite sets*, Studies in Combinatorics, G.-C. Rota, ed., Mathematical Association of America, Washington, DC, 1978, pp. 22-79.

[7] G. O. H. KATONA, *Intersection theorems for finite sets*, Acta Math. Acad. Sci. Hungar., 15 (1964), pp. 329-337.

[8] ———, *A simple proof of the Erdös-Ko-Rado theorem*, J. Combin. Theory (B), 13 (1972), pp. 183-184.

[9] D. J. KLEITMAN, *On a combinatorial conjecture of Erdös*, J. Combin. Theory, 1 (1966), pp. 209-214.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

010

ON THE REDUCTION OF A MATRIX TO TRIANGULAR OR DIAGONAL FORM BY CONSIMILARITY*

YOO PYO HONG† AND ROGER A. HORN‡

Abstract. We study the problem of reducing a given n-by-n complex matrix A to triangular or diagonal form by a transformation of the form $A \rightarrow SA\bar{S}^{-1}$, where S is a nonsingular n-by-n complex matrix. We also consider the special case of this reduction in which S is unitary, and a generalization to the problem of simultaneously reducing a family of matrices in this way. Natural analogues of eigenvalues and eigenvectors arise in this context; they have both familiar and unfamiliar properties.

AMS(MOS) subject classifications. 15A21, 15A23

1. Introduction. In the theory of univalent complex analytic functions in the unit disc, an important role is played by quadratic inequalities of the form

$$x^*Ax \geqq |x^TBx| \quad \text{for all } x \in \mathbb{C}^n, \tag{1.1}$$

where A and B are n-by-n complex matrices, A is Hermitian and positive semidefinite, and B is symmetric [1]. Under a nonsingular change of variables $x \rightarrow Sy$, the matrices transform according to the laws $A \rightarrow S^*AS$, $B \rightarrow S^TBS$, and it is easy to show that if A is nonsingular (and hence is positive definite), there is always a nonsingular S that transforms A and B simultaneously into diagonal form in this way.

Now suppose that A and B are n-by-n matrices, not necessarily related by (1.1), with B symmetric and A Hermitian and nonsingular, but not necessarily definite. If there exists a nonsingular S such that both $S^*AS = \Lambda$ and $S^TBS = M$ are diagonal, then $A^{-1}\bar{B} = (S\Lambda^{-1}S^*)((S^*)^{-1}\bar{M}\bar{S}^{-1}) = S(\Lambda^{-1}\bar{M})\bar{S}^{-1}$, i.e., $A^{-1}\bar{B}$ has the property that there is a nonsingular R such that $R(A^{-1}\bar{B})\bar{R}^{-1}$ is diagonal. This necessary condition is also sufficient to ensure that A and B can be reduced simultaneously to diagonal form by these mixed congruences [3]. It is an example of how the notions of consimilarity and condiagonalizability arise naturally.

A second example is an old, and often rediscovered, result about complex symmetric matrices. If A is an n-by-n complex symmetric matrix, there is a unitary matrix U and a nonnegative diagonal matrix Σ such that $UAU^T = \Sigma$. This may be thought of as a theorem about diagonalization by unitary congruence, or it may be thought of as a singular value decomposition, but if we write it as $UA\bar{U}^{-1} = \Sigma$, we see that it is of the same form as the first example, but with a unitary consimilarity matrix $R = U$.

In the next section, we introduce the basic concepts involved with the theory of consimilarity. In the third section, we treat the problem of reducing a given matrix to upper triangular from by consimilarity, and in the last section we consider reduction to diagonal form by consimilarity.

2. Basic notions. We denote by M_n the set of n-by-n complex matrices. Two matrices $A, B \in M_n$ are said to be *consimilar* if there is a nonsingular $R \in M_n$ such that $A = RB\bar{R}^{-1}$. Like ordinary similarity, consimilarity is an equivalence relation on M_n, and we may ask which equivalence classes contain triangular or diagonal representatives. A matrix $A \in M_n$ is said to be *contriangularizable* if there exists a nonsingular

* Received by the editors June 28, 1984, and in final revised form November 7, 1984.

† Department of Mathematics and Computer Science, University of Maryland, Baltimore County, Catonsville, Maryland 21228.

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

$R \in M_n$ such that $R^{-1}A\bar{R}$ is upper triangular; it is said to be *condiagonalizable* if R can be chosen so that $R^{-1}A\bar{R}$ is diagonal.

If $A \in M_n$ is condiagonalizable and $R^{-1}A\bar{R} = \Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$, then $A\bar{R} = R\Lambda$. If $R = (r_1, \cdots, r_n)$ with each $r_i \in \mathbb{C}^n$, this identity says that $A\bar{r}_i = \lambda_i r_i$ for $i = 1, 2, \cdots, n$. A nonzero vector x such that $A\bar{x} = \lambda x$ is said to be a *coneigenvector* of A; the scalar λ is a *coneigenvalue* of A. The identity $A\bar{R} = R\Lambda$ says that every nonzero column of the matrix R is a coneigenvector of A. Since the columns of R are independent if and only if R is nonsingular, we see that a matrix $A \in M_n$ is condiagonalizable if and only if it has n independent coneigenvectors. To this extent, the theory of condiagonalization is entirely analogous to the theory of ordinary diagonalization.

But every matrix has at least one eigenvalue, and it has only finitely many distinct eigenvalues; in this regard, the theory of coneigenvalues is rather different. If $A\bar{x} = \lambda x$, then $e^{-i\theta}A\bar{x} = A(\overline{e^{i\theta}x}) = e^{-i\theta}\lambda x = (e^{-2i\theta}\lambda)(e^{i\theta}x)$ for all $\theta \in \mathbb{R}$. Thus, if λ is a coneigenvalue of A, then so is $e^{i\theta}\lambda$ for all $\theta \in \mathbb{R}$. On the other hand, if $A\bar{x} = \lambda x$, then $A\bar{A}x = A(\overline{A\bar{x}}) = A(\overline{\lambda x}) = \bar{\lambda}A\bar{x} = \bar{\lambda}\lambda x = |\lambda|^2 x$, so a scalar λ is a coneigenvalue of A only if $|\lambda|^2$ is an eigenvalue of $A\bar{A}$. The example $A = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, for which $A\bar{A} = -2I$ has no nonnegative eigenvalues, shows that there are matrices that have no coneigenvalues at all; iA is an example of a 2-by-2 Hermitian matrix with no coneigenvalues. It is known, however, that if $A \in M_n$ and n is odd, then A must have at least one coneigenvalue [4], a result analogous to the fact that every real matrix of odd order has at least one real eigenvalue.

Thus, in contrast to the theory of ordinary eigenvalues, a matrix may have infinitely many distinct coneigenvalues or it may have no coneigenvalues at all. If a matrix has a coneigenvalue, it is sometimes convenient to select from among the coneigenvalues of equal modulus the unique nonnegative one as a representative.

The necessary condition we have just observed for the existence of a coneigenvalue is also sufficient.

PROPOSITION 2.1. *Let $A \in M_n$ and let $\lambda \geqq 0$ be given. Then λ is an eigenvalue of $A\bar{A}$ if and only if $+\sqrt{\lambda}$ is a coneigenvalue of A.*

Proof. If $\lambda \geqq 0$, $\sqrt{\lambda} \geqq 0$, and $A\bar{x} = \sqrt{\lambda}x$ for some $x \neq 0$, then $A\bar{A}x = A(\overline{A\bar{x}}) = A(\overline{\sqrt{\lambda}\,x}) = \sqrt{\lambda}\ A\bar{x} = \sqrt{\lambda}\ \sqrt{\lambda}\ x = \lambda x$.

Conversely, if $A\bar{A}x = \lambda x$ for some $x \neq 0$, there are two possibilities:

(a) $A\bar{x}$ and x are dependent, or

(b) $A\bar{x}$ and x are independent.

In the former case, there is some $\mu \in \mathbb{C}$ such that $A\bar{x} = \mu x$, which says that μ is a coneigenvalue of A. But then $\lambda x = A\bar{A}x = A(\overline{A\bar{x}}) = A(\overline{\mu x}) = \bar{\mu}A\bar{x} = \bar{\mu}\mu x = |\mu|^2 x$, so $|\mu| = +\sqrt{\lambda}$. Since $e^{-2i\theta}\mu$ is a coneigenvalue associated with the coneigenvector $e^{i\theta}x$ for any $\theta \in \mathbb{R}$, we conclude that $+\sqrt{\lambda}$ is a coneigenvalue of A. Notice that $A\bar{A}(A\bar{x}) = A(\overline{A\bar{A}x}) = A(\overline{\lambda x}) = \lambda(A\bar{x})$ and $A\bar{A}x = \lambda x$, so if λ is a simple nonnegative eigenvalue of $A\bar{A}$, (a) must always be the case.

In the latter case (b) (which could occur if λ is a multiple eigenvalue of $A\bar{A}$), the vector $y = A\bar{x} + \sqrt{\lambda}\,x$ is nonzero and is a coneigenvector corresponding to the coneigenvalue $+\sqrt{\lambda}$ since $A\bar{y} = A\bar{A}x + \sqrt{\lambda}\,A\bar{x} = \lambda x + \sqrt{\lambda}\,A\bar{x} = \sqrt{\lambda}(A\bar{x} + \sqrt{\lambda}\,x) = \sqrt{\lambda}\,y$. □

We have seen that to each distinct nonnegative eigenvalue of $A\bar{A}$ there corresponds a coneigenvector of A, a result analogous to a familiar fact in the ordinary theory of eigenvectors. The following result extends this analogy a bit further.

PROPOSITION 2.2. *Let $A \in M_n$ be given, and let $x_1, x_2, \cdots, x_k$ be coneigenvectors of A with corresponding coneigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_k$. If $|\lambda_i| \neq |\lambda_j|$ whenever $1 \leqq i, j \leqq k$ and $i \neq j$, then $\{x_1, \cdots, x_k\}$ is an independent set.*

Proof. Each x_i is an eigenvector of $A\bar{A}$ with associated eigenvalue $|\lambda_i|^2$. The vectors $x_1, \cdots, x_k$ are independent because they are eigenvectors of the matrix $A\bar{A}$ and their associated eigenvalues $|\lambda_1|^2, \cdots, |\lambda_k|^2$ are distinct by assumption. □

This result, together with Proposition 2.1, gives a lower bound on the number of independent coneigenvectors of a given matrix, and yields a sufficient condition for condiagonalizability that is analogous to a familiar sufficient condition for ordinary diagonalizability. We give a more general condition in Theorem 4.3.

COROLLARY 2.3. *Let $A \in M_n$ be given. If $A\bar{A}$ has k distinct nonnegative eigenvalues, then A has at least k independent coneigenvectors. If $k = n$, A is condiagonalizable. If $k = 0$, A has no coneigenvectors at all.*

These bounds on the number of independent coneigenvectors are sharp. For $A = J_n$, an elementary Jordan block

$$J_n = \begin{bmatrix} 1 & 1 & & 0 \\ & 1 & \ddots & \\ & & \ddots & 1 \\ 0 & & & 1 \end{bmatrix} \in M_n,$$

$A\bar{A} = J_n^2$ has 1 as its only nonnegative eigenvalue. The coneigenvector equation $J_n\bar{x} = x$ is easily seen to have only real solutions, so every coneigenvector is also an eigenvector, and the subspace of eigenvectors is one-dimensional. Direct sums of elementary Jordan blocks can therefore be used to give examples of matrices $A \in M_n$ such that $A\bar{A}$ has k distinct nonnegative eigenvalues and A has exactly k independent coneigenvectors, for any $n \geqq k \geqq 1$.

For a given matrix, the set of coneigenvectors corresponding to a given coneigenvalue is not generally a subspace of $\mathbb{C}^n$ over $\mathbb{C}$, but is a subspace of $\mathbb{C}^n$ over $\mathbb{R}$.

The notion of consimilarity can be generalized by replacing the complex field with an arbitrary field F and replacing the operation of complex conjugation by an automorphism on F [7, p. 27].

3. Contriangularization and unitary condiagonalization. Any complex matrix can be reduced to upper triangular form by a unitary similarity, but an analogous reduction is not always possible for consimilarity. If $A \in M_n$ is given, and if there is a nonsingular $S \in M_n$ such that $A = S\Delta\bar{S}^{-1}$ for some upper triangular $\Delta \in M_n$, then $A\bar{A} = S\Delta\bar{S}^{-1}\bar{S}\bar{\Delta}S^{-1} = S(\Delta\bar{\Delta})S^{-1}$, and hence $A\bar{A}$ is similar to $\Delta\bar{\Delta}$. But $\Delta\bar{\Delta}$ has nonnegative real main diagonal entries, so a necessary condition for a given matrix A to be contriangularizable is that all the eigenvalues of $A\bar{A}$ are nonnegative. This condition is also sufficient to ensure that the contriangularization can be accomplished with a unitary transformation. Given Proposition 2.1, the proof is similar to the proof of Schur's ordinary unitary triangularization theorem.

THEOREM 3.1. *Let $A \in M_n$ be given. There exists a unitary $U \in M_n$ and an upper triangular $\Delta \in M_n$ such that $A = U\Delta U^T$ if and only if all the eigenvalues of $A\bar{A}$ are real and nonnegative. Under this condition, the main diagonal entries of Δ may be chosen to be nonnegative.*

Proof. The necessity of the eigenvalue condition has already been shown. To show that the condition is sufficient, assume that all the eigenvalues of $A\bar{A}$ are nonnegative, and let λ be an eigenvalue of $A\bar{A}$. By Proposition 2.1, there is a coneigenvector v of A corresponding to the coneigenvalue $+\sqrt{\lambda}$. Since v/v^*v is also a coneigenvector, there is no loss of generality if we assume that v is a unit vector and $A\bar{v} = +\sqrt{\lambda}\, v$.

unitary matrix that has these vectors as its respective columns. The first column of the matrix $\bar{V}_1^T A \bar{V}_1$ has entries $v_i^* A \bar{v} = +\sqrt{\lambda}\, v_i^* v = +\sqrt{\lambda}\, \delta_{i1}$ because of orthonormality and the relation $A\bar{v} = +\sqrt{\lambda}\, v$. Thus, all but perhaps the first of the entries in the first column of $\bar{V}_1^T A \bar{V}_1$ must be zero. If we write this matrix in partitioned form as

$$\bar{V}_1^T A \bar{V}_1 = \left[\begin{array}{c|c} +\sqrt{\lambda} & w^T \\ \hline 0 & \\ \vdots & A_2 \\ 0 & \end{array}\right], \quad w \in \mathbb{C}^{n-1}, \quad A_2 \in M_{n-1}, \tag{3.2}$$

we see that

$$(\bar{V}_1^T A \bar{V}_1)(\overline{\bar{V}_1^T A \bar{V}_1}) = V_1^* A \bar{A} V_1 = \left[\begin{array}{c|c} \lambda & +\sqrt{\lambda}\, w^T + w^T A_2 \\ \hline 0 & \\ \vdots & A_2 \bar{A}_2 \\ 0 & \end{array}\right].$$

The eigenvalues of $A\bar{A}$ (all nonnegative by assumption) are therefore λ together with the eigenvalues of $A_2\bar{A}_2$. Thus, the matrix $A_2 \in M_{n-1}$ obtained by this process of reduction also has the property that all the eigenvalues of $A_2\bar{A}_2$ are nonnegative.

The process of reduction can now be repeated with A_2 and its successors at most $n-1$ times to obtain

$$\bar{V}_{n-1}^T \cdots \bar{V}_2^T \bar{V}_1^T A \bar{V}_1 \bar{V}_2 \cdots \bar{V}_{n-1} = \begin{bmatrix} \sigma_1 & & * \\ & \ddots & \\ 0 & & \sigma_n \end{bmatrix} = \Delta,$$

where each V_i is unitary and Δ is upper triangular with nonnegative main diagonal entries σ_i. If we set $U = V_1 V_2 \cdots V_{n-1}$, we have $A = U \Delta U^T$ as desired. □

Not every matrix $A \in M_n$ has the property that all the eigenvalues of $A\bar{A}$ are nonnegative, but Hermitian positive semidefinite matrices and symmetric matrices do have this property. If $A \in M_n$ is Hermitian and positive definite (nonsingular), then it has a Hermitian positive definite square root $A^{1/2}$ and $A\bar{A}$ is similar to $A^{-1/2}(A\bar{A})A^{1/2} = A^{1/2}\bar{A}A^{1/2}$, which is positive definite (and hence has positive eigenvalues) because it is congruent to the positive definite matrix $\bar{A}$. A limiting argument now shows that if $A \in M_n$ is Hermitian and positive semidefinite, then all the eigenvalues of $A\bar{A}$ are nonnegative and hence A is unitarily contriangularizable. The example $A = \left(\begin{smallmatrix} 0 & i \\ -i & 0 \end{smallmatrix}\right)$ shows that it is not sufficient to assume that A is merely Hermitian. If A is complex symmetric, however, it is always unitarily contriangularizable because $A\bar{A} = A\bar{A}^T = AA^*$, and AA^* is Hermitian and positive semidefinite for any $A \in M_n$. But if A is symmetric and $A = U\Delta U^T$ for some unitary U, then $\Delta = U^* A \bar{U} = U^* A^T \bar{U} = (U^* A \bar{U})^T = \Delta^T$, so Δ must be symmetric, too. Since a symmetric triangular matrix must be diagonal, we conclude that every symmetric complex matrix is unitarily condiagonalizable.

COROLLARY 3.3. *A matrix $A \in M_n$ is symmetric if and only if there are a unitary $U \in M_n$ and a nonnegative diagonal $\Sigma \in M_n$ such that $A = U\Sigma U^T$.*

This result is often attributed to Schur [8], but earlier proofs were offered by Hua [5], Seigel [9], and Jacobson [6]; historical priority must apparently be given to Takagi [10]. In the setting of consimilarity, complex symmetric matrices are analogous to normal matrices in the sense that complex symmetric matrices can be reduced to diagonal form by unitary consimilarity and normal matrices can be reduced to diagonal form by unitary similarity. Corollary 3.3 may be thought of as an analogue for consimilarity of the spectral theorem for normal matrices.

Although our proof of Corollary 3.3 is completely elementary, it may be useful to have another elementary proof that proceeds directly to the diagonalization without first proving the triangularization Theorem 3.1. If $A \in M_n$ is a given symmetric matrix, then $A\bar{A} = AA^*$ is Hermitian and hence $AA^* = V\Lambda V^*$ for some unitary $V \in M_n$ and a real, (in fact, nonnegative) diagonal $\Lambda \in M_n$. Notice that the matrix $B \equiv V^*A\bar{V}$ is also symmetric, and $B\bar{B} = V^*A\bar{V}V^T\bar{A}V = V^*AA^*V = \Lambda$ is real. If we denote the real and imaginary parts of B by B_1 and B_2, respectively, then $B = B_1 + iB_2$, B_1 and B_2 are real symmetric matrices, and $B\bar{B} = (B_1^2 + B_2^2) - i(B_1B_2 - B_2B_1) = \Lambda$, so $B_1B_2 - B_2B_1 = 0$, i.e., B_1 and B_2 are commuting real symmetric matrices. There is, therefore, a real orthogonal $Q \in M_n$ such that $B_1 = Q\Lambda_1Q^T$ and $B_2 = Q\Lambda_2Q^T$, with Λ_1 and Λ_2 both real diagonal. But then $B = B_1 + iB_2 = Q(\Lambda_1 + i\Lambda_2)Q^T = V^*A\bar{V}$, so $A = (VQ)\Lambda(VQ)^T = W\Lambda W^T$ with a unitary $W = VQ$ and a diagonal $\Lambda = \Lambda_1 + i\Lambda_2$. This is almost the factorization in Corollary 3.3, and the argument is completed by observing that $\Lambda = \Sigma D^2 = D\Sigma D$ with $\Sigma = |\Lambda|$ a nonnegative diagonal matrix and D a diagonal matrix with main diagonal entries with unit modulus. Then D is unitary and $A = W\Lambda W^T = WD\Sigma DW^T = U\Sigma U^T$ with $U = WD$. The heart of this argument is due to Siegel [9], but it seems to be little-known. The same sort of argument can be used to deduce a normal form for a complex skew-symmetric matrix under unitary consimilarity.

Symmetry is a necessary and sufficient condition for unitary condiagonalization and is a sufficient, but not necessary, condition for condiagonalizability. We consider necessary and sufficient conditions for general nonsingular condiagonalizability in the next section.

One might be interested in conditions under which a family $\mathcal{F} = \{A_i \colon i \in \mathcal{I}\} \subset M_n$ of complex symmetric matrices is simultaneously unitarily condiagonalizable, i.e., there is a single unitary $U \in M_n$ such that UA_iU^T is diagonal for all $i \in \mathcal{I}$. A necessary and sufficient condition is that $A_i\bar{A}_j = A_j\bar{A}_i$ for all $i, j \in \mathcal{I}$, i.e., each product $A_i\bar{A}_j$ is Hermitian. This result and more general results about simultaneous unitary contriangularization of a family of matrices may be found in [2].

4. Condiagonalization. Our objective is to give a simple condition for a given matrix to be condiagonalizable, and as a first step we prove the following lemma. The motivation for this result is that if a given matrix $A \in M_n$ is consimilar to a scalar matrix, then $A = S(\lambda I)\bar{S}^{-1} = \lambda S\bar{S}^{-1}$ and $A\bar{A} = \lambda S\bar{S}^{-1}\bar{\lambda}\bar{S}S^{-1} = |\lambda|^2 I$. Matrices with this property (that $A\bar{A}$ is a scalar matrix) are the basic building blocks from which condiagonalizable matrices are constructed.

LEMMA 4.1. *A matrix $A \in M_n$ has the property that $A\bar{A} = I$ if and only if there exists a nonsingular $S \in M_n$ such that $A = S\bar{S}^{-1}$.*

Proof. We have just seen that the stated condition is necessary. To show that it is sufficient, define $S_\theta \equiv e^{i\theta}A + e^{-i\theta}I$ for any $\theta \in \mathbb{R}$ and observe that

$$(4.2) \qquad A\bar{S}_\theta = A(e^{-i\theta}\bar{A} + e^{i\theta}I) = e^{-i\theta}A\bar{A} + e^{i\theta}A = e^{i\theta}A + e^{-i\theta}I = S_\theta.$$

Since A has only finitely many eigenvalues, there is some $\theta_0 \in \mathbb{R}$ such that $-e^{-2i\theta_0}$ is not an eigenvalue of A. For this value of θ, $S_{\theta_0} = e^{i\theta_0}(A + e^{-2i\theta_0}I)$ is nonsingular and $A = S_{\theta_0}\bar{S}_{\theta_0}^{-1}$ from (4.2). □

We can now state and prove a necessary and sufficient condition for condiagonalizability.

THEOREM 4.3. *Let $A \in M_n$. There exists a nonsingular $S \in M_n$ and a diagonal $\Lambda \in M_n$ such that $A = S\Lambda\bar{S}^{-1}$ if and only if $A\bar{A}$ is a diagonalizable matrix with real nonnegative eigenvalues and* rank A = rank $A\bar{A}$.

Proof. The stated conditions are clearly necessary since $A\bar{A} = S\Lambda\bar{S}^{-1}\bar{S}\bar{\Lambda}S^{-1} = S|\Lambda|^2S^{-1}$ and the rank of both $A\bar{A}$ and A is the number of nonzero diagonal entries in Λ. Conversely, if $A\bar{A}$ is diagonalizable and has nonnegative eigenvalues there is a nonsingular $S \in M_n$ and a nonnegative diagonal $\Lambda \in M_n$ such that $A\bar{A} = S\Lambda S^{-1}$. There is no loss of generality to assume that like diagonal entries in Λ are grouped together and that $\Lambda = \lambda_1 I_{n_1} \oplus \lambda_2 I_{n_2} \oplus \cdots \oplus \lambda_k I_{n_k}$, where $I_{n_i} \in M_{n_i}$ and $\lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_k \geqq 0$. We then have

$$S^{-1}A\bar{A}S = S^{-1}A\bar{S}\bar{S}^{-1}\bar{A}S = (S^{-1}A\bar{S})(\overline{S^{-1}A\bar{S}}) = \Lambda.$$

If we set $B = S^{-1}A\bar{S}$, then (since consimilarity is an equivalence relation) it will suffice to show that B is condiagonalizable if $B\bar{B} = \Lambda$. Since Λ is real, $\Lambda = \bar{\Lambda} = (\overline{B\bar{B}}) = \bar{B}B = B\bar{B}$, so B and $\bar{B}$ commute. Thus, $B\Lambda = B(B\bar{B}) = BB\bar{B} = (B\bar{B})B = \Lambda B$, so B and Λ also commute. If we write B in block form as

$$B = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1k} \\ \vdots & B_{22} & & \vdots \\ & & \ddots & \\ B_{k1} & \cdots & \cdots & B_{kk} \end{bmatrix}$$

with block sizes conformal to those of

$$\Lambda = \begin{bmatrix} \lambda_1 I_{n_1} & & 0 \\ & \ddots & \\ 0 & & \lambda_k I_{n_k} \end{bmatrix}, \quad I_{n_i} \in M_{n_i}, \quad i = 1, 2, \cdots, k,$$

then the equation $B\Lambda = \Lambda B$ says that $\lambda_i B_{ij} = \lambda_j B_{ij}$ for all $i = 1, 2, \cdots, k$. Since $\lambda_i \neq \lambda_j$ if $i \neq j$, we conclude that $B_{ij} = 0$ if $i \neq j$ and hence B is block diagonal

$$B = \begin{bmatrix} B_{11} & & 0 \\ & \ddots & \\ 0 & & B_{kk} \end{bmatrix}$$

with diagonal blocks the same size as those of Λ. The equation $B\bar{B} = \Lambda$ means that each $B_{ii}\bar{B}_{ii} = \lambda_i I$ for $i = 1, 2, \cdots, k$. Notice that B_{ii} must be nonsingular if $\lambda_i > 0$. If $\lambda_i > 0$ we can write this equation as

$$\left[\frac{1}{\sqrt{\lambda_i}} B_{ii}\right]\left[\overline{\frac{1}{\sqrt{\lambda_i}} B_{ii}}\right] = I_{n_i}$$

and we can use Lemma 4.1 to conclude that there is a nonsingular $S_i \in M_{n_i}$ such that $B_{ii} = S_i(\sqrt{\lambda_i}\, I_{n_i})\bar{S}_i^{-1}$. If $\lambda_k = 0$, then

$$\begin{aligned} \operatorname{rank}(B_{11}) + \operatorname{rank}(B_{22}) + \cdots + \operatorname{rank}(B_{kk}) &= \operatorname{rank}(B) \\ = \operatorname{rank}(A) = \operatorname{rank}(A\bar{A}) &= \operatorname{rank}(\Lambda) = n_1 + n_2 + \cdots + n_{k-1}. \end{aligned}$$

This means that the rank of $B_{k,k}$ is zero, so the last block B_{kk} must actually be a zero block if $\lambda_k = 0$. In this event, we can write $0 = B_{kk} = S_k(\sqrt{\lambda_k} I_{n_k})\bar{S}_k^{-1}$, where $S_k \in M_{n_k}$ is an arbitrary nonsingular matrix. If we set $S = S_1 \oplus \cdots \oplus S_k$, we have shown in all cases that $B = S(\sqrt{\lambda_1}\, I_{n_1} \oplus \cdots \oplus \sqrt{\lambda_k}\, I_{n_k})\bar{S}^{-1}$ and we are done. □

The special case in which A is a complex symmetric matrix is handled easily by the theorem, since $A\bar{A} = AA^*$ is Hermitian in this case and hence is diagonalizable.

Moreover, rank $A = \operatorname{rank} AA^*$ for any $A \in M_n$, so the hypotheses of the theorem are satisfied whenever A is a symmetric matrix. The theorem implies that every symmetric matrix can be condiagonalized but does not yield directly the fact that the condiagonalization can be accomplished with a unitary transformation.

If $A \in M_n$ is Hermitian and positive definite, then $A\bar{A}$ is similar to the Hermitian positive definite matrix $A^{1/2}\bar{A}A^{1/2}$ and is therefore diagonalizable and has positive eigenvalues. Since $\operatorname{rank}(A) = \operatorname{rank}(A\bar{A}) = n$ in this case, the theorem guarantees that every Hermitian positive definite matrix is condiagonalizable. The condiagonalization can be accomplished with a complex orthogonal transformation, though this does not follow directly from the theorem. A Hermitian positive semidefinite matrix need not be condiagonalizable, as the example $A = \begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix}$ shows; $\operatorname{rank}(A) = 1 \neq \operatorname{rank}(A\bar{A}) = 0$.

Theorem 4.3 is a special case of a general theorem about consimilarity: Two matrices $A, B \in M_n$ are consimilar if and only if (a) $A\bar{A}$ is similar to $B\bar{B}$, and (b) rank $(A) = $ rank (B), rank $(A\bar{A}) = $ rank $(B\bar{B})$, rank $(A\bar{A}A) = $ rank $(B\bar{B}B)$, $\cdots$, and so on, for all n such alternating products with at most n terms. Using this characterization of consimilar matrices, one can deduce the following for every $A \in M_n$: A is consimilar to a real matrix, A is consimilar to A^T, $\bar{A}$, and A^*, and A is consimilar to a Hermitian matrix [4].

We have a simple necessary and sufficient condition for a single matrix to be condiagonalizable, but what about simultaneous condiagonalization of more than one matrix?

If a given family $\mathcal{F} = \{A_i \colon i \in \mathcal{I}\} \subset M_n$ has the property that there is a nonsingular $S \in M_n$ such that $A_i = S\Lambda_i\bar{S}^{-1}$ and Λ_i is diagonal for all $i \in \mathcal{I}$, then each $A_i\bar{A}_j = S\Lambda_i\bar{S}^{-1}\bar{S}\bar{\Lambda}_jS^{-1} = S\Lambda_i\bar{\Lambda}_jS^{-1}$. The family $\mathcal{G} = \{A_i\bar{A}_j \colon i, j \in \mathcal{I}\}$ is, therefore, a commuting family of diagonalizable matrices. Moreover, $A_i\bar{A}_j + A_j\bar{A}_i = S(\Lambda_i\bar{\Lambda}_j + \Lambda_j\bar{\Lambda}_i)S^{-1} = S(2 \operatorname{Re}(\Lambda_i\bar{\Lambda}_j))S^{-1}$ has only real eigenvalues and $A_i\bar{A}_j - A_j\bar{A}_i = S(2i \operatorname{Im}(\Lambda_i\bar{\Lambda}_j))S^{-1}$ has only imaginary eigenvalues. These necessary conditions are also sufficient.

THEOREM 4.4. *Let $\mathcal{F} = \{A_i \colon i \in \mathcal{I}\} \subset M_n$ and $\mathcal{G} = \{A_i\bar{A}_j \colon i, j \in \mathcal{I}\}$. There exists a nonsingular $S \in M_n$ such that $SA_i\bar{S}^{-1}$ is diagonal for all $i \in \mathcal{I}$ if and only if*

(a) *A_i is condiagonalizable for all $i \in \mathcal{I}$, i.e., for each $i \in \mathcal{I}$*

(a1) *$A_i\bar{A}_i$ is diagonalizable,*

(a2) *all the eigenvalues of $A_i\bar{A}_i$ are nonnegative, and*

(a3) rank $(A_i) = $ rank $(A_i\bar{A}_i)$;

(b) *$A_i\bar{A}_j$ is diagonalizable for all $i, j \in \mathcal{I}$;*

(c) *$\mathcal{G}$ is a commuting family; and*

(d) *for all $i, j \in \mathcal{I}$, $A_i\bar{A}_j + A_j\bar{A}_i$ has only real eigenvalues and $A_i\bar{A}_j - A_j\bar{A}_i$ has only imaginary eigenvalues.*

Proof. The necessity of these four conditions is easily verified; we proceed to show that they are sufficient. Conditions (b) and (c) imply that $\mathcal{G}$ is simultaneously diagonalizable, so there exists a nonsingular $S \in M_n$ such that $S^{-1}A_i\bar{A}_jS = \Lambda_{ij}$ is diagonal for all $i, j \in \mathcal{I}$. Condition (d) implies that $\Lambda_{ij} = \bar{\Lambda}_{ji}$. If we set $B_i \equiv S^{-1}A_i\bar{S}$, then B_i is consimilar to A_i and the family $\{B_i \colon i \in \mathcal{I}\}$ has the properties

(a′) B_i is condiagonalizable for all $i \in \mathcal{I}$,

(b′) $B_i\bar{B}_j = S^{-1}A_i\bar{S}\bar{S}^{-1}\bar{A}_jS = \Lambda_{ij}$ is diagonal for all $i, j \in \mathcal{I}$,

(c′) $\Lambda_{ij} = \bar{\Lambda}_{ji}$ for all $i, j \in \mathcal{I}$.

Moreover, the family $\{B_i\}$ is simultaneously condiagonalizable if and only if the family $\{A_i\}$ is simultaneously condiagonalizable.

Since each B_i is condiagonalizable, we know that $B_i\bar{B}_i = \Lambda_{ii}$ has only nonnegative diagonal entries and that rank $(B_i) = \operatorname{rank}(\Lambda_{ii})$. If all $\Lambda_{ii} = 0$, then all $B_i = 0$ and we are done. If some $\Lambda_{00} \neq 0$, there is a (real) permutation matrix P such that

$$(4.5) \qquad P\Lambda_{00}P^{-1} = \begin{bmatrix} \lambda_1 I_1 & & & 0 \\ & \lambda_2 I_2 & & \\ & & \ddots & \\ 0 & & & \lambda_k I_k \end{bmatrix}, \qquad I_j \in M_{n_j}$$

$n_1 + n_2 + \cdots + n_k = n$, $\lambda_1 > \lambda_2 > \cdots > \lambda_{k-1} > \lambda_k \geqq 0$, for some k with $1 \leqq k \leqq n$. Since P is real, we have $P\Lambda_{ij}P^{-1} = PB_i\bar{B}_j\bar{P}^{-1} = (PB_i\bar{P}^{-1})(\bar{P}\bar{B}\bar{P}^{-1})$ for all $i, j \in \mathscr{I}$. Observe that $P\Lambda_{ij}P^{-1}$ is diagonal, $P\Lambda_{ij}P^{-1} = \bar{P}\bar{\Lambda}_{ji}\bar{P}^{-1}$, and $PB_i\bar{P}^{-1}$ is consimilar to B_i. Thus, there is no loss of generality to assume that Λ_{00} has the block form (4.5) as a direct sum of k distinct nonnegative scalar matrices.

Notice that $B_0\Lambda_{00} = B_0\bar{\Lambda}_{00} = B_0\bar{B}_0B_0 = \Lambda_{00}B_0$, so Λ_{00} commutes with B_0. Partition each B_i conformally with (4.5) so that

$$(4.6) \qquad B_i = \begin{bmatrix} B_i^{11} & \cdots & B_i^{1k} \\ \vdots & \ddots & \vdots \\ B_i^{k1} & \cdots & B_i^{kk} \end{bmatrix}, \qquad B_i^{ij} \in M_{n_j}, \quad j = 1, 2, \cdots, k.$$

By equating the i, j blocks of both sides of the identity $B_0\Lambda_{00} = \Lambda_{00}B_0$, we obtain the identities $B_0^{ij}\lambda_j I_j = \lambda_i I_i B_0^{ij}$, or $(\lambda_j - \lambda_i)B_0^{ij} = 0$. Since $\lambda_i \neq \lambda_j$ if $i \neq j$, $B_0^{ij} = 0$ if $i \neq j$, and B_0 has the block diagonal form

$$(4.7) \qquad B_0 = \begin{bmatrix} B_0^{11} & & 0 \\ & \ddots & \\ 0 & & B_0^{kk} \end{bmatrix}, \qquad B_0^{ii} \in M_{n_i}.$$

Since B_0 is condiagonalizable, rank $(B_0) =$ rank $(B_0\bar{B}_0) =$ rank Λ_{00}. If $\lambda_k = 0$, then $B_0^{kk} = 0$. If $\lambda_k > 0$, then B_0^{kk} is nonsingular. In either event, B_0^{ii} is nonsingular for all $i = 1, 2, \cdots, k-1$. Now let $l \in \mathscr{I}$ and equate the corresponding i, j blocks of the identity $B_0\bar{B}_l = \Lambda_{0l}$. We find $B_0^{ii}\bar{B}_1^{ij} = 0$ if $i \neq j$ for all $l \in \mathscr{I}$. Since each B_0^{ii} is nonsingular for $i = 1, 2, \cdots, k-1$, this implies that $B_l^{ij} = 0$ for all $j \neq i$, $i = 1, 2, \cdots, k-1$. By applying the same reasoning to the last block row of the product $B_l\bar{B}_0 = \bar{\Lambda}_{0l}$, we find that $B_l^{kj}\bar{B}_0^{jj} = 0$ for all $j = 1, 2, \cdots, k$ and hence $B_l^{kj} = 0$, $j = 1, 2, \cdots, k-1$. Thus, every B_l is in block diagonal form with the block structure of (4.5) and (4.7).

If $k = n$, i.e., B_0 has n distinct coneigenvalues, our argument shows that every B_l is diagonal and we are done. If $k < n$, there is at least one scalar matrix of size two or greater among the diagonal blocks of Λ_{00}, and we must make a further argument. Because every B_l is a direct sum of smaller matrices in the same way, it suffices to consider the case in which $k = 1$, i.e., $\Lambda_{00} = \lambda_{00}I$. Since we are assuming $\Lambda_{00} \neq 0$, $\lambda_{00} > 0$ and B_0 is nonsingular. Consider $B_0\bar{B}_l = \Lambda_{0l}$ for $l \neq 0$, and suppose not all the matrices Λ_{0l} are scalar matrices. If Λ_{01} is not a scalar matrix, then it can be put into block scalar diagonal form like (4.5) by a permutation similarity which, since it is real, is also a consimilarity. Now apply the same permutation (con)similarity to every B_l. Although B_0 may be altered by this transformation, the scalar matrix $B_0\bar{B}_0 = \Lambda_{00}$ is not altered. But $\Lambda_{01}B_0 = B_0\bar{B}_1B_0 = B_0\bar{\Lambda}_{10} = B_0\Lambda_{01}$, so B_0 commutes with Λ_{01} and hence B_0 is block diagonal with the same block structure as Λ_{01}, and all the diagonal blocks of B_0 are nonsingular. If we examine the i, j blocks of the identity $B_0\bar{B}_l = \Lambda_{0l}$ as before, we find that each B_l has the same block diagonal structure as Λ_{01} and B_0. For each diagonal block of Λ_{01}, look at the corresponding diagonal block of each Λ_{0l}. If they are all scalar matrices, stop. If any one is not a scalar matrix, then focus on that block, permute, and argue again that the resulting sub-blocks are found in all the B_i's.

In at most $n-1$ steps, this process of successive refinement into diagonal blocks results in a new family of matrices $\{C_i\}$ in which each C_i is consimilar to B_i (and hence to the original A_i) by a single permutation (con)similarity. Moreover, each C_i has the same block diagonal structure

$$C_i = \begin{bmatrix} C_i^{(1)} & & 0 \\ & \ddots & \\ 0 & & C_i^{(m)} \end{bmatrix}, \qquad C_i^{(j)} \in M_{n_i}, \quad i \in \mathscr{I},$$

$n_1+n_2+\cdots+n_m=n$, and $C_0^{(j)}\bar{C}_i^{(j)}=\lambda_i^{(j)}I_j$, $I_j\in M_{n_j}$. Each $C_0^{(j)}$ is condiagonalizable, so there is a nonsingular $S_j\in M_{n_j}$ such that $S_jC_0^{(j)}\bar{S}_j^{-1}=\lambda_jI_j$ with $\lambda_j>0$. Then $C_0^{(j)}\bar{C}_i^{(j)}=S_j^{-1}\lambda_jI_j\bar{S}_j\bar{C}_i^{(j)}=\lambda_i^{(j)}I_j$, and $\bar{S}_j\bar{C}_i^{(j)}S_j^{-1}=(\lambda_i^{(j)}/\lambda_j)S_jI_jS_j^{-1}=(\lambda_i^{(j)}/\lambda_i)I_j$. This shows that the matrix $S_1\oplus\cdots\oplus S_m$ simultaneously condiagonalizes every C_i.

REFERENCES

[1] H. GRUNSKY, *Koeffizientenbedingungen für schlicht abbildende meromorphe Funktionen*, Math. Z., 45 (1939), pp. 29-61.

[2] Y. P. HONG AND R. A. HORN, *On simultaneous reduction of families of matrices to triangular or diagonal form by unitary congruences*, Linear Multilinear Algebra, to appear.

[3] Y. P. HONG, R. A. HORN AND C. R. JOHNSON, *On the reduction of pairs of Hermitian or symmetric matrices to diagonal form by congruence*, Linear Algebra Appl., to appear.

[4] Y. P. HONG AND R. A. HORN, *A canonical form for matrices under consimilarity*, Technical Report 415, Dept. Mathematical Sciences, The Johns Hopkins University, Baltimore, MD, October 30, 1984.

[5] L. K. HUA, *On the theory of automorphic functions of a matrix variable* I*—geometric basis*, Amer. J. Math., 66 (1944), pp. 470-488.

[6] N. JACOBSON, *Normal semi-linear transformations*, Amer. J. Math., 61 (1939), pp. 45-58.

[7] ———, *The Theory of Rings*, American Mathematical Society, New York, 1943.

[8] I. SCHUR, *Ein Satz Über quadratische Formen mit komplexen Koeffizienten*, Amer. J. Math., 67 (1945), pp. 472-480.

[9] C. L. SIEGEL, *Symplectic geometry*, Amer. J. Math., 65 (1943), pp. 1-86.

[10] T. TAKAGI, *On an algebraic problem related to an analytic theorem of Caratheodory and Fejer and on an allied theorem of Landau*, Japanese J. Math., 1 (1927), pp. 83-93.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

011

SUPER LINE-CONNECTIVITY PROPERTIES OF CIRCULANT GRAPHS*

F. T. BOESCH† AND J. F. WANG‡

Abstract. The connection between line-connectivity concepts of graphs and indices of network reliability is well-known. Of particular interest in such studies are the circulant graphs because the connected ones have the largest possible value of line-connectivity λ of p-point, degree r, regular graphs, namely $\lambda = r$. In this work, we define the higher order line-connectivity measure N_i as the number of line-disconnecting sets of order i. Regular degree r, p-point graphs having $\lambda = r$ satisfy $N_\lambda \geqq p$. Such graphs which attain this lower bound are called super-λ. In this work we determine the necessary and sufficient conditions for a circulant to be super-λ. In addition we determine a lower bound on N_i for $\lambda \leqq i \leqq 2r-3$. It is shown that a special class of circulants, known as Harary graphs, achieve this lower bound for all these values of i.

Key words. circulant, connecting, disconnecting line set, edge connectivity, line connectivity, λ-graph, Harary graph, super line connectivity, super edge connectivity, vulnerability

AMS(MOS) subject classifications. 05C40, 68E10, 94C15

Introduction. To study the vulnerability of a communication network it is customary to represent the network by an undirected graph. In this work we consider certain graph theoretic optimization problems related to the design of invulnerable networks. The terminology and notations of the graphs follow the book by Harary [7]. In this graph model one usually assumes that the graph of the network is connected and the network is said to have failed if the graph becomes disconnected when a set of lines called a *disconnecting line set* fails. A measure of the vulnerability of a graph to line failure is the *line-connectivity* λ which is the minimum order of a disconnecting line set.

Suppose now that the cost of building a network is proportional to the number of lines employed. Then the following optimization problem describes the design of invulnerable graphs:

Determine the minimum number of lines q among all
graphs having $\lambda \geqq n$ for given values of p and n.

It is easily verified that $q \geqq \lceil np/2 \rceil$ where $\lceil x \rceil$ denotes the smallest integer not less than x. To verify that this lower bound is in fact the solution to the stated optimization problem, it suffices to show that for any p and n ($p \geqq n+1$) there exist graphs having $\lceil np/2 \rceil$ lines and $\lambda = n$. The solution was originally given by Harary [8] who constructed a special class of graphs having these properties. However there are many graphs which achieve this optimal, and we refer to them as *λ-graphs.* In fact when either p or n is even, the necessary and sufficient condition for a p-point graph to be a λ graph is that it be regular of degree $\delta = \lambda = n$. Examples abound which show that an arbitrary regular graph need not be a λ-graph. However many classes of regular graphs are λ-graphs. Hence one might wish to impose further constraints to enable a comparison of the vulnerability of λ-graphs. An obvious possibility results from the observation that some λ-graphs have the property that removing a minimum disconnecting line set may divide the graph into two parts having $p/2$ points each, while other λ-graphs

* Received by the editors May 27, 1983 and in revised form November 12, 1984. This work was supported under National Science Foundation grant ECS-8100652.

† Stevens Institute of Technology, Hoboken, New Jersey 07030.

‡ National Cheng-Kung University, Tainan, Taiwan, Republic of China.

can only have a single point isolated by the removal of a minimum disconnecting line set. This motivates the following definition.

DEFINITION 1. If δ denotes the minimum degree of any point in a graph G, then G is said to be super-λ if every disconnecting line set of order λ is the incidence set of a point of degree δ.

We note that if G is a regular graph of degree δ on p points, then a super-λ graph has the maximum possible value of λ for given values of p and δ, namely $\lambda = \delta$. However, the example $K_2 \times C_3$ serves to verify that $\lambda = \delta$ is not sufficient to insure that a graph is super-λ.

We now turn our attention to a special class of regular graphs which includes those shown to be λ-graphs by Harary. In order to define them, we assume that the points of a graph are labelled 0, 1, 2, $\cdots, p-1$, and we refer to point i instead of saying the point labelled i.

DEFINITION 2. The circulant graph $C_p(n_1, n_2, \cdots, n_k)$ or briefly $C_p(n_i)$ where $0 < n_1 < \cdots < n_k < (p+1)/2$ has $i \pm n_1, i \pm n_2, \cdots, i \pm n_k (\text{mod } p)$ adjacent to each point i. The sequence (n_i) is called the *jump sequence* and the n_i are called the *jumps*. The earliest connectivity result for circulants is due to Harary [8] who showed that $C_p(1, 2, \cdots, k)$, which we call *Harary graphs*, has both point and line-connectivity equal to δ. A generalization of the line connectivity property of circulants is given in [2] where it is shown that the circulant $C_p(n_1, n_2, \cdots, n_k)$ is super-λ if $n_1 = 1$ and $k \geqq 2$. However, it may be noted that these conditions are not necessary for a circulant to be super-λ as shown by $C_{10}(2, 5)$ which is super-λ but not even isomorphic to a circulant having a jump of unity.

Herein we determine the necessary and sufficient conditions for a circulant to be super-λ. We then turn our attention to the problem of determining N_i, the number of disconnecting line sets of order i (where $i > \lambda$) for the Harary graphs.

The reason for considering such numbers is that the problem of finding the probability of disconnection for a network having equal and independent line failures can be reduced to finding all the N_i values of the corresponding graph. A complete discussion of the connection between these two problems is given in [3]. Here we merely note that in order to minimize this probability of disconnection over all p-point, regular degree δ graphs, one must first maximize λ and then minimize all the N_i. Graphs which are super-λ are of interest because a regular degree δ graph G with $\lambda = \delta$ has $N_\lambda \geqq p$, with equality achieved if and only if G is super-λ. Further discussions of these reliability problems can be found in [4], [5].

The super-λ class of circulants. Clearly $C_p(n_1)$ has $\lambda = 0$, or 2, but it is never super-λ. We proceed to establish a sequence of lemmas that will determine when any circulant is super-λ. However we need two preliminary results. Theorem 1 follows immediately from number theory. Theorem 2, which apparently has never been stated in English, is due to Mader [9]. His result applies to the class of point-symmetric graphs, which include all circulants.

THEOREM 1. *The circulant $C_p(n_1, n_2, \cdots, n_k)$ is connected if and only if* $\gcd(p, n_1, n_2, \cdots, n_k) = 1$.

THEOREM 2 (Mader [9]). *Every connected, point-symmetric graph has $\lambda = \delta$.*

LEMMA 1. *Let $n \geqq 3$, $m \geqq 3$ and G be the union of m point-disjoint cycles of length n, say $C_0, C_1, \cdots, C_{m-1}$, together with n-independent lines between C_i and C_{i+1}*(mod m) *for all i, $0 \leqq i \leqq m-1$. Then G is super-λ.*

Proof. Note that G is regular of degree 4. One possible structure for G with $m = 3$ and $n = 3$ is illustrated in Fig. 1.

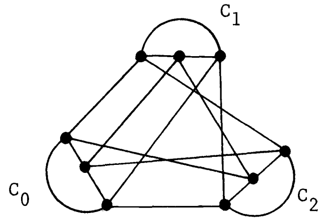


FIG. 1

Let U be a minimum order disconnecting line set of G. First it is claimed that since C_i is a cycle, U must contain at least two lines from C_i for some i. Otherwise, each C_i is connected and can be coalesced into a single point. In this case U must disconnect an n multiple line cycle C_m, or $|U| \geqq 2n > 4 = \delta(G)$, which is impossible. Now let these two lines be $e_1 = \{x_1, y_1\}$, $e_2 = \{x_2, y_2\}$. Suppose e_1 and e_2 are not adjacent. Without loss of generality, it can be assumed that x_1, x_2 are in one component and y_1, y_2 are in the other component.

From Fig. 2 it is clear that there are two line-disjoint paths from x_1, x_2 through C_{i+1} to y_1, y_2 and two line-disjoint paths from x_1, x_2 through C_{i-1} to y_1, y_2. Consequently, there are four line-disjoint paths from x_1, x_2 through C_{i+1} or C_{i-1} to y_1, y_2. Thus $|U| \geqq 2+4 > 4 = \delta(G)$, a contradiction.

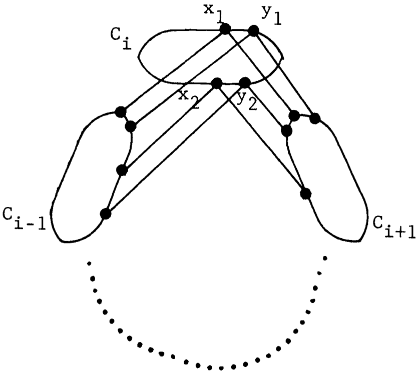


FIG. 2

Hence e_1 and e_2 have to be adjacent to one point w, and so $e_1 = \{x_1, w\}$, $e_2 = \{x_2, w\}$ (see Fig. 3). It is now claimed that w should be isolated by U. Suppose not. Without loss of generality, it can be assumed that there exists another point w' in C_{i+1} such that w' and w are adjacent. And it follows that x_1, x_2 are in one component and w', w are in the other component. But (see Fig. 3) there exist two line-disjoint paths from

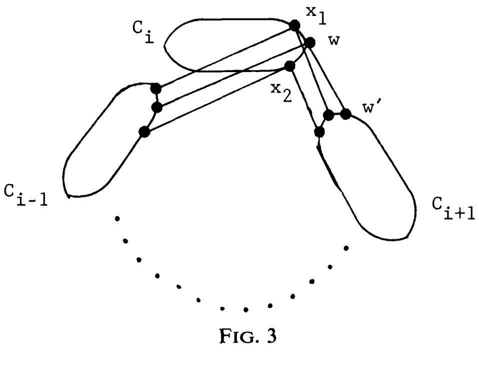


FIG. 3

x_1, x_2 through C_{i+1} to w' and there exists a third path, line-disjoint from the other 2 paths, that connects x_1, x_2 through C_{i-1} to w. Thus $|U| \geqq 2+1+2=5>4$, a contradiction.

Therefore w is isolated by U and the result is established. □

LEMMA 2. *If $G=C_p(n_1, n_2)$ is connected, then G is super-λ for $n_2<p/2$.*

Proof. If $\max\{\gcd(p, n_1), \gcd(p, n_2)\} \leqq 2$ then since G is connected, it follows that $\gcd(p, n_1)$ and $\gcd(p, n_2)$ cannot be equal to two simultaneously. Therefore, $\min\{\gcd(p, n_1), \gcd(p, n_2)\}=1$. Hence there is some number r such that either rn_1 or $rn_2 = 1 (\bmod p)$, where $\gcd(r, p)=1$.

Assume without loss of generality that $rn_1 = 1(\bmod p)$ and that $rn_2(\bmod p) = x(2 \leqq x \leqq p-1)$. Let a denote the minimum of x and $p-x$.

It follows from the work of Ádám [1] that $C_p(n_1, n_2)$ is isomorphic to $C_p(1, a)$. Thus by the theorem of Bauer, Boesch, Suffel and Tindell [2], G is super-λ.

Now if $\max\{\gcd(p, n_1), \gcd(p, n_2)\}=m \geqq 3$, then G can be viewed as the union of m point-disjoint cycles of length p/m together with p/m independent lines between C_i and $C_{i+1}(\bmod m)$ for all $i, 0 \leqq i \leqq m-1$ (see Fig. 4). Therefore, by Lemma 1, G is super-λ. □

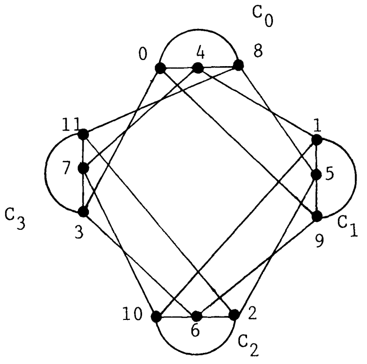


FIG. 4. $C_{12}(3, 4)$, $m=4$.

LEMMA 3. *Let $G=C_p(n_1, n_2, \cdots, n_k)$ be connected, $k \geqq 2$, $n_k<p/2$. Then G is super-λ.*

Proof. Let U be a minimum disconnecting line set. The basic step for a proof by induction on k is provided by Lemma 2. Assume $k \geqq 3$ and Lemma 3 holds for the circulant graph with fewer than k jumps. Obviously U must contain at least one line from $C_p(n_i)$ for some $i, 1 \leqq i \leqq k$. Since $C_p(n_i)$ is a line-disjoint union of cycles, U contains at least two lines from $C_p(n_i)$. This is true because if $x=\{u, v\} \in U$ then u and v should be in different components. But u and v are in the same cycle, thus at least one other line in this cycle should be removed. Now let $\hat{U}=\{e \mid e \text{ in } U \text{ but not in } C_p(n_i)\}$ and $G'=G-C_p(n_i)$. Thus $G'=G-C_p(n_i)=C_p(n_1, n_2, \cdots, n_{i-1}, n_{i+1}, \cdots, n_k)$. There are two cases.

Case 1. G' is disconnected. By Theorem 1, $\gcd(p, n_1, n_2, \cdots, n_{i-1}, n_{i+1}, \cdots, n_k)=m \geqq 2$, and since G is connected, $\gcd(n_i, m)=1$. In this case G' consists of m components C_j. Each C_j is isomorphic to

$$G_1 = C_{p/m}\left(\frac{n_1}{m}, \cdots, \frac{n_{i-1}}{m}, \frac{n_{i+1}}{m}, \cdots, \frac{n_k}{m}\right)$$

(see Fig. 5) and there are p/m independent lines between C_j and C_{j+1} for all $j, 0 \leqq j \leqq m-1$, if $m \geqq 3$. There are $2p/m$ lines when $m=2$. By the induction hypothesis, G_1 is

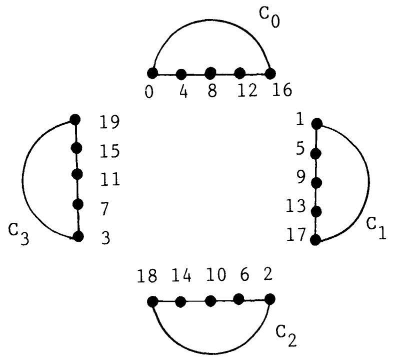


FIG. 5. $C_{20}(3,4)$, $G'=G-C_{20}(3)$.

super-λ and so are all C_j. It is claimed that $\hat{U}$ must disconnect at least one C_j for some j, $0\leqq j\leqq m-1$. Otherwise, since $G-U$ is disconnected,

$$|U|=2\frac{p}{m}>2\frac{2n_k}{m}\geqq 4(k-1)=4k-4\geqq 2k,$$

which is impossible.

Now we know $\hat{U}$ disconnects at least one C_j and since C_j is super-λ, $|\hat{U}|\geqq\lambda(C_j)=\delta(C_j)=2k-2$. But $|\hat{U}|\leqq|U|-2=2k-2$, thus $|\hat{U}|=|U|-2=\lambda(C_j)=2k-2$. Hence it is concluded that $\hat{U}$ is a minimum line disconnecting set for C_j. Therefore $\hat{U}$ isolates a point (say w) in C_j. Since the number of independent lines joining C_j and C_{j+1} is $p/m>2$ if $m\geqq 3$, and is p if $m=2$, and since $|U|=|\hat{U}|+2$, each point in $V(C_j)-\{w\}$ is connected to C_{j-1} and C_{j+1} in $G-U$. It follows that all points in G except w are in the same component of $G-U$. Therefore G is super-λ.

Case 2. G' is connected. By induction hypothesis, G' is super-λ and since $\hat{U}$ is a disconnecting set in G', $|\hat{U}|\geqq\lambda(G')=2k-2$. But $|\hat{U}|\leqq|U|-2=2k-2$, so $|\hat{U}|=|U|-2=2k-2$. Thus $\hat{U}$ is a minimum disconnecting line set. Hence $\hat{U}$ isolates a point (say w) in G', and so all the points in $V(G')-\{w\}$ are in the same component of $G'-\hat{U}$. Since U is in the union of $\hat{U}$ and two lines not in G', U cannot disconnect $G'-w$. It follows that all the points in $V(G)-\{w\}=V(G')-\{w\}$ are in the same component. Thus G is super-λ. □

LEMMA 4. *Let $G=C_p(n_1,n_2,\cdots,n_k)$ be connected, with $k\geqq 2$, p even, and $n_k=p/2$. Then G is not super-λ if and only if*

$$k=\frac{p/2+1}{2}\quad \textit{and}\quad n_i=2i\quad \textit{for all } i,\ 1\leqq i\leqq k-1.$$

Proof. Let p be even, $n_k=p/2$, $k=(p/2+1)/2$ and $n_i=2i$ for all i, $1\leqq i\leqq k-1$. Let U be the set of all the lines joining point i and point $p/2+i$ for all i, $0\leqq i\leqq p/2-1$. Then $G-U=C_p(2,4,6,8,\cdots,2(k-1))$ becomes disconnected. Each component of $G-U$ has $p/2$ points. Since G has $\lambda=\delta$,

$$\lambda(G)=2k-1=2\left(\frac{p/2+1}{2}\right)-1=\frac{p}{2}.$$

But $|U|=p/2$, thus U is a minimum disconnecting line set for G. Now as U does not isolate a point, it follows that G is not super-λ.

It remains to show that if $k\neq(p/2+1)/2$ then G is super-λ, and if $k=(p/2+1)/2$ but $n_i=2i$ for all i, $1\leqq i\leqq k-1$, is not true, then G is also super-λ.

Case 1. $k \neq (p/2+1)/2$. Let U be a minimum disconnecting line set for G. First it is claimed that at least one line of U is in $C_p(p/2)$. Suppose not, and let G_1 be the multi-graph obtained by coalescing points $i<p/2$ and $i+p/2$ into a single point i of G_1. There is one line between points i and j (where $i, j<p/2$) in G_1 for each of the following possible lines of G:

$$\{i,j\}, \quad \{i, j+p/2\}, \quad \{i+p/2, j\}, \quad \{i+p/2, j+p/2\}.$$

It follows that G contains either 0, 2, or 4 of these lines for each $i, j<p/2$. Thus there are an even number of lines between each pair of points of G_1, and $\lambda(G_1)$ will be even. However it follows from the path version of the min cut-max flow theorem that $\lambda(G) \leqq \lambda(G_1)$. But by assumption there was a minimum disconnecting line set of G that did not contain any of the coalesced lines which implies $\lambda(G)=\lambda(G_1)$. This is a contradiction as $\lambda(G)$ is the odd number $2k-1$.

Now let $\hat{U}=\{e \mid e$ in U but not in $C_p(p/2)\}$ and $G_2=G-C_p(p/2)$. Obviously $|\hat{U}| \leqq |U|-1=2k-2$.

There are two subcases.

Subcase 1. G_2 is disconnected. If G_2 is disconnected then by Theorem 1, $\gcd(p, n_1, n_2, \cdots, n_{k-1})=m \geqq 2$. And since G is connected, $\gcd(p/2, m)=1$. Now as p is divisible by m, let $p=jm$. If j were even, then m would be a factor of $p/2$, which contradicts $p/2$ and m being relatively prime. Now as p is even and j is odd, $m=2s$ for some s. Hence $\gcd(p/2, m)=\gcd(js, 2s)=s=1$. It follows that $m=2$ and $p/2$ is odd. Moreover, each component of G_2 is isomorphic to (see Case 1 of Lemma 3)

$$G_3=C_{p/2}\left(\frac{n_1}{2}, \frac{n_2}{2}, \cdots, \frac{n_{k-1}}{2}\right) \quad \text{and} \quad \frac{n_{k-1}}{2}<\frac{p/2}{2}.$$

By Lemma 3, each component of G_2 is super-λ. Since $k \neq (p/2+1)/2$ and since $n_k=p/2>n_{k-1} \geqq 2(k-1)$, it follows that $2k-1<p/2$. Now if $\hat{U}$ does not disconnect one component of G_2, then in order to disconnect G, U must consist of all the lines in $C_p(p/2)$. This implies $|U|=p/2>2k-1=\delta(G)$, a contradiction. So $|\hat{U}| \geqq \lambda(G_3)=2k-2$, and since $|\hat{U}| \leqq |U|-1=2k-2$, it is concluded that $|\hat{U}|=2k-2=|U|-1$. Therefore $\hat{U}$ is a minimum disconnecting line set for a component of G_2. It follows that $\hat{U}$ isolates a point (say w) in this component. Now we know U is the union of $\hat{U}$ and one line (say e) in $C_p(p/2)$. If e is not adjacent to w then $G-U$ is still connected, a contradiction. (See Fig. 6). Thus U isolates a point in G, and so G is super-λ.

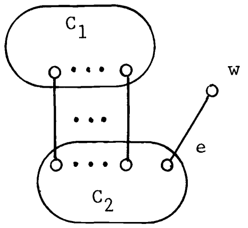


FIG. 6. $G_2-\hat{U}=C_1 \cup C_2 \cup \{w\}$, $G_2=G-C_p(p/2)$.

Subcase 2. G_2 is connected. If G_2 is connected, then $\hat{U}$ must disconnect G_2. Suppose not, then

$$\begin{aligned} G-U &= [G_2 \cup C_p(p/2)]-[\hat{U} \cup \{\text{some lines in } C_p(p/2)\}] \\ &= [G_2-\hat{U}] \cup \{\text{some lines in } C_p(p/2)\}, \end{aligned}$$

i.e., $G-U$ is the union of connected graph $(G_2-\hat{U})$ and some other line set. This means that $G-U$ is still connected, a contradiction.

Then using arguments similar to Subcase 1, we obtain that $\hat{U}$ is a minimum disconnecting line set for G_2. Consequently, U isolates a point in G, and G is super-λ.

Case 2. $k=(p/2+1)/2$ but $n_i=2i$ for all i, $1 \leqq i \leqq k-1$ is not true.

In this case, we first show that $G_2=G-C_p(p/2)$ is always connected. Suppose G_2 is disconnected. Then from the proof of Subcase 1, gcd $(p, n_1, n_2, \cdots, n_{k-1})=m=2$.

It follows that $n_i \geqq 2i$ for all i, $1 \leqq i \leqq k-1$. But by the assumption $k=(p/2+1)/2$, i.e., $2(k-1)=(p/2)-1=n_k-1 \geqq n_{k-1}$. Thus $n_{k-1} \leqq 2(k-1)$, and since all n_i are even, $1 \leqq i \leqq k-1$, it follows that $n_{k-2} \leqq n_{k-1}-2=2(k-2)$, $n_{k-3} \leqq 2(k-3) \cdots$ etc. Finally we obtain $n_i \leqq 2i$ for all i, $1 \leqq i \leqq k-1$. Thus we have $2i \leqq n_i \leqq 2i$ for all i, $1 \leqq i \leqq 2k-1$, a contradiction.

Now we know G_2 is connected. Then following the proof in Subcase 2, G is super-λ. □

THEOREM 3. *A connected circulant is super-λ unless it is $C_p(a)$ or $C_{2n}(2, 4, 6, \cdots, n-1, n)$ for n odd.*

Proof. The theorem follows immediately from Lemmas 3 and 4. □

Higher order line-connectivity measures. We define N_i as the total number of disconnecting line sets of order i. In general it is difficult to determine all the N_i values of a graph; in fact Provan and Ball [10] have shown that the problem is NP-hard. However, for the class of circulants known as Harary graphs we shall evaluate many of the N_i explicitly.

THEOREM 4. *Let $G=C_p(1, 2, \cdots, k)$, $2 \leqq k < p/2$, and U be a disconnecting line set. If $|U|=i$ and $\lambda \leqq i \leqq 4k-3$ then U isolates exactly one point and*

$$N_i=\binom{q-2k}{i-2k}p,$$

where q is the number of lines in a p-point graph G, and N_i is the number of disconnecting line sets of order i.

Proof. Let the p points of G be labeled as $0, 1, \cdots, p-1$, and let C be any component of $G-U$. We shall assume $|V(C)|=m \geqq 2$ for all such C. Now let C be decomposed into n contiguous parts (a part of G whose points are labeled contiguous) say $C_1, C_2, \cdots, C_n$ and let $|V(C_i)|=m_i$ for all $1 \leqq i \leqq n$, and the gaps between C_i and $C_{(i+1) \bmod n}$ are denoted by g_i. Assume the C_i are maximal.

First it is claimed that $m \geqq k+1$. Otherwise, if $m \leqq k$ then every point in C has degree at most $m-1$. Thus $|U| \geqq (2k-m+1)m=2km-m^2+m$, and let $b=|U|-(4k-3) \geqq 2km-m^2+m-4k+3$; then it can be easily verified that

if $m=2$ then $b \geqq 1>0$;

if $m=3$ then $b \geqq 2k-3>0$;

if $m \geqq 4$ then $b \geqq m(k-m)+(m-4)k+m+3>0$.

All the above show $|U|>4k-3$, a contradiction.

So we know that the points of any component of $G-U$ are at least $k+1$ in number, if that component has two or more points.

Second, it is claimed that not all the gaps are trivial, i.e., at least one g_i will have at least two points. Suppose not; then every gap only has a single point and note the union of all gaps $(g_i \cup g_2 \cup \cdots \cup g_n)$ are also components of $G-U$, and this union of gaps must also have at least $k+1$ points, i.e., $n \geqq k+1$. Then it can be shown that the number of lines from each g_i to C, is not smaller than 2 if $k=2$, and the number of lines from each g_i to C, is at least 4, if $k \geqq 3$. So

$$|U| \geqq 2n \geqq 2k+2=6>5=4k-3 \quad \text{if } k=2,$$

$$|U| \geqq 4n \geqq 4k+4>4k-3 \quad \text{if } k \geqq 3.$$

All the above are contradictions.

Thus without loss of generality, it can be assumed that g_1 has two end points labeled x and y. Let the points in C adjacent to x and y be labeled as u and v, respectively. (See Fig. 7).

Consider now the following definitions:

Since the points of a circulant are labelled 0 to $p-1$ we assume they are located in clockwise increasing order on a circle.

Let h_{x1}, h_{x2} denote the number of lines from x to C in the clockwise and counterclockwise direction, respectively.

Let h_{y1}, h_{y2} denote the number of lines from y to C in the clockwise and counterclockwise direction, respectively.

Let h_{u1}, h_{u2} denote the number of lines from u to $G-C$ in the clockwise and counterclockwise direction, respectively.

Let h_{v1}, h_{v2} denote the number of lines from v to $G-C$ in the clockwise and counterclockwise direction, respectively.

Then $|U| \geqq h_{x1}+h_{x2}+h_{y1}+h_{y2}+h_{u1}+h_{u2}+h_{v1}+h_{v2}-d$ where d is the number of lines which overlap in the above count.

Note $d \leqq 4$, and 4 is the worst case when the lines $\{u, x\}$, $\{u, y\}$, $\{v, x\}$, $\{v, y\}$ all overlap.

In order to determine the value of d, there are three subcases discussed below:

Case 1. (see Fig. 8). If there exist some C_i for $i \geqq 3$, then there are at least two lines from C_i to $G-C-\{x\}-\{y\}$. The two lines are $\{m, n\}$ and $\{0, p\}$.

Case 2. (see Fig. 9). If there are no C_i for $i \geqq 3$, then $n=1$ or 2. When n is one, then $|V(g_1)| \geqq k+1$ (since g_1 becomes a component) then the lines $\{u, y\}$, $\{v, x\}$, are never counted, and there are two lines $\{u, x\}$, $\{v, y\}$, which are counted twice.

Case 3. (see Fig. 10). The remaining case is $n=2$ and there is another gap g_2 between C_1 and C_2. If $|V(C_1)|+|V(C_2)| \geqq 4$ then there are at least two lines from g_2 to $C-\{u\}-\{v\}$.

If $|V(C_1)|+|V(C_2)|=3$ then $k=2$, and counting the number of lines from C to $G-C$, we obtain $|U|>5=4k-3$.

Now note $k-h_{x2}$ is the number of lines from x to $G-C$ in the counterclockwise direction, since the jump sizes are contiguous; so $h_{u2} \geqq k-h_{x2} \rightarrow h_{u2}+h_{x2} \geqq k$. And $k-h_{u1}$ is the number of lines from u to C in the clockwise direction, again because the jump sizes are contiguous, $h_{x1} \geqq k-h_{u1} \rightarrow h_{x1}+h_{u1} \geqq k$.

Similarly, $h_{v1}+h_{y1} \geqq k$, $h_{v2}+h_{y2} \geqq k$. Therefore we have $|U| \geqq 4k-d$. And in Case 1, $|U| \geqq 4k-4+2=4k-2$, in Case 2, $|U| \geqq 4k-2$, and in Case 3, $|U| \geqq 4k-3$. All are contradictions.

This completes the proof that every component of $G-U$ has two or more points is impossible; hence U isolates a point.

Now suppose U isolates two or more points, say t points; if $t \leqq 2k$ then $|U| \geqq 2kt-t(t-1)/2$. Let $r=|U|-4k+3$ and then it can be verified that

if $t=2$ then $r \geqq 2>0$;

if $t=3$ then $r \geqq 2k>0$;

if $t \geqq 4$ then $r>0$;

if $t \geqq 2k+1$ then $|U| \geqq 2kt-2kt/2=kt \geqq 2k^2+k>4k-3$.

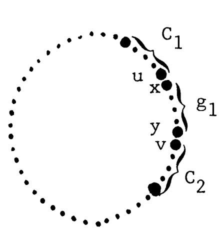


FIG. 7

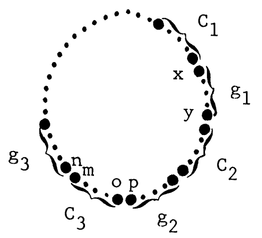


FIG. 8

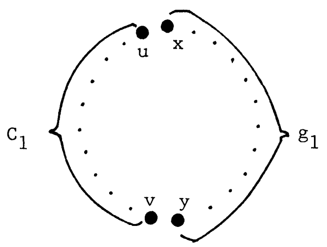


FIG. 9

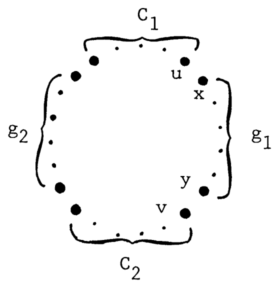


FIG. 10

All the above are contradictions, so $t=1$. Now let U_1 and U_2 be two disconnecting line sets, and $\lambda \leqq |U_1| = |U_2| \leqq 4k-3$. If U_1 isolates point x, U_2 isolates point $y \neq x$, then $U_1 \neq U_2$. Otherwise $U_1 = U_2$ will isolate more than one point, which is impossible. Thus we have

$$N_i = \binom{q-2k}{i-2k} p, \quad \text{and} \quad 2k = \lambda \leqq i \leqq 4k-3. \qquad \square$$

Conclusions. In conclusion we note that any regular degree $2k$ graph on p-points will have

$$N_i \geqq \binom{q-2k}{i-2k} p$$

as this lower bound counts only those order i disconnecting line sets which are obtained from line incidence sets at points. Hence we have extended the result in [2] to show that Harary graphs not only minimize N_λ but all N_i for $\lambda \leqq i \leqq 4k-3$. For i larger than this, the analysis used in the proof of Theorem 4 cannot be used. In fact, it can be shown that the Harary graphs do *not* minimize all the N_i; for example C_8 $(1,3)$ has $N_9 = \binom{16}{7} - 4{,}096$ while C_8 $(1,2)$ has $N_9 = \binom{16}{7} - 3{,}528$.

Finally we note that Theorem 3 may be viewed as giving the two forbidden circulants in the class of all super-λ graphs which are circulants. However, these two graphs do not characterize the class of nonsuper-λ, point-symmetric graphs. Namely $K_2 \times K_4$ is point-symmetric but not super-λ. However, it is not isomorphic to either of the two forbidden circulants; in fact $K_2 \times K_4$ is not isomorphic to any circulant.

REFERENCES

[1] A. ÁDÁM, *Research problem* 2-10, J. Combin. Theory, 3 (1967), p. 393.

[2] D. BAUER, F. BOESCH, C. SUFFEL AND R. TINDELL, *Connectivity extremal problems and the design*

of reliable probabilistic networks, The Theory and Application of Graphs, G. Chartrand, Y. Alavi, D. Goldsmith, L. Lesniak-Foster and D. Lick, eds., Wiley, New York, 1981, pp. 45-54.

[3] ———, *Combinatorial optimization problems in the analysis and design of probabilistic networks*, to appear.

[4] F. BOESCH, *Introduction to basic network problems*, in The Mathematics of Networks, S. Burr, ed., Proceedings of Symposia in Applied Mathematics 26, American Mathematical Society, Providence, RI, 1982, pp. 1-29.

[5] ———, *An introduction to the theory of large-scale networks*, Large-Scale Networks: Theory and Design, F, Boesch, ed., IEEE Press, New York, 1976, pp. 3-10.

[6] F. BOESCH AND R. TINDELL, *Circulants and their connectivities*, J. Graph Theory, 8 (1984), pp. 487-499.

[7] F. HARARY, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.

[8] ———, *The maximum connectivity of a graph*, Proc. National Academy of Sciences, USA, 48 (1962), pp. 1142-1146.

[9] W. MADER, *Minimale n-fach kantenzusammenhängende Graphen*, Math. Ann., 191 (1971), pp. 21-28.

[10] J. PROVAN AND M. BALL, *The complexity of counting cuts and computing the probability that a graph is connected*, Management Science and Statistics Report 81-002, Univ. Maryland, College Park, 1981.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

012

VECTOR COMPUTATIONS FOR SPARSE LINEAR SYSTEMS*

DAVID R. KINCAID†, THOMAS C. OPPE† AND DAVID M. YOUNG†

Abstract. We are interested in the development of algorithms, based on iterative methods, and software for the solution of large sparse systems of linear algebraic equations with emphasis on systems arising in the numerical solution of partial differential equations. The objective is to develop algorithms and software which are effective when used with a vector computer such as the Control Data CYBER 205 or the CRAY 1. A package of programs, known as ITPACK, has been developed for use on conventional, or *scalar* machines. A number of "short-range" modifications to ITPACK, including changes in the data storage format and changes in the programming, but not in the algorithms used, have been made and tested on a number of numerical examples. Preliminary work is described on "long-range" modifications which will involve extensive changes in the basic algorithms in order to achieve efficient vectorization.

AMS(MOS) subject classification. 65F10

1. Introduction. The advent of high-performance vector computers such as the Control Data CYBER 205 and the CRAY 1 is having a profound effect on the areas of numerical analysis and mathematical software. This is true, in particular, for iterative algorithms and software for solving large sparse systems of linear algebraic equations. While there is a large potential gain achievable for many problems by using a vector computer as compared with using a scalar computer, nevertheless this potential gain can often only be realized by a careful choice of algorithms. It is often the case that an iterative algorithm which is effective when used with a conventional or *scalar* computer may not be as effective as expected on a vector machine. At the same time, an algorithm which is very inefficient for a scalar machine may turn out to be surprisingly efficient when used with a vector machine.

Normally, only a small part of the potential gain in using a vector computer can be realized by making a direct conversion from a scalar program to a vector program. Sometimes, however, it is possible to realize a substantial portion of the potential improvement by making "short-range" modifications to a program such as, for example, changing the data structure and the programming but not changing the basic algorithm. In many cases, however, a complete restructuring of the entire computer program, including both the algorithm and the programming, is needed before the true potential of a vector computer can be achieved.

In this paper, we describe some of our work on the development of iterative algorithms and software which are designed to be effective when used on vector computers. As our starting point, we consider a package of subroutines, known as ITPACK 2C, which we developed for solving sparse linear systems by a variety of iterative methods. (See Kincaid, Respess, Young and Grimes [1982].) This package, which is described briefly in § 2, has been developed for a scalar computer. In § 3, we discuss the iterative algorithms currently included in ITPACK from the standpoint of vectorization. In § 4, we describe short-range modifications of ITPACK. These modifications, which primarily involve the use of a different storage scheme have been incorporated into a new package, ITPACKV 2C (Kincaid, Oppe, Respess, and Young [1984]). The result of numerical experiments and the effectiveness of various changes

* Received by the editors February 16, 1984, and in revised form October 15, 1984. This work was supported in part by the National Science Foundation under grant MCS-8214731, by the U.S. Department of Energy under grant DE-A505-81ER10954, by the North Atlantic Treaty Organization under grant 648183, and by Control Data Corporation, under grant 84PCR54B with the University of Texas at Austin.

† Center for Numerical Analysis, University of Texas at Austin, Austin, Texas 78712.

to the package are also given. In § 5, we give a brief description of our work on long-range modifications.

2. The ITPACK package. The ITPACK software package has been developed over a period of several years at the Center for Numerical Analysis of the University of Texas at Austin. The package provides for the iterative solution of the linear system

$$Au = b, \tag{2.1}$$

where A is a given $N \times N$ matrix, b is a given $N \times 1$ column vector and the $N \times 1$ column vector u is to be determined. The matrix A is assumed to be nonsingular and sparse. While the routines of ITPACK 2C often work in more general cases, they are primarily designed to handle cases where A is symmetric and positive definite.

The ITPACK package provides for the solution of (2.1) by any one of seven alternative iterative algorithms. Each algorithm involves a *basic* iterative method and, except for one algorithm, an acceleration procedure. Each basic iterative method has the form

$$u^{(n+1)} = Gu^{(n)} + k,$$

where for some nonsingular matrix Q we have $G = I - Q^{-1}A$ and $k = Q^{-1}b$. The basic iterative algorithms used in ITPACK include the Jacobi method, the successive overrelaxation (SOR) method, the symmetric SOR (SSOR) method, and the RS method. The RS method is applicable to the case where the matrix A is a *red-black* matrix of the form

$$A = \begin{pmatrix} D_R & H \\ K & D_B \end{pmatrix}, \tag{2.2}$$

where D_R and D_B are square diagonal matrices. If we write (2.1) in the form

$$\begin{pmatrix} D_R & H \\ K & D_B \end{pmatrix} \begin{pmatrix} u_R \\ u_B \end{pmatrix} = \begin{pmatrix} b_R \\ b_B \end{pmatrix}, \tag{2.3}$$

then the *reduced system* is

$$(D_B - KD_R^{-1}H)u_B = b_B - KD_R^{-1}b_R.$$

Consequently, the RS method is defined by

$$u_B^{(n+1)} = (D_B^{-1}KD_R^{-1}H)u_B^{(n)} + D_B^{-1}b_B - D_B^{-1}KD_R^{-1}b_R.$$

The acceleration procedures used in ITPACK include Chebyshev acceleration and conjugate gradient acceleration. In each case the procedure is defined by

$$u^{(n+1)} = \rho_{n+1}\{u^{(n)} + \gamma_{n+1}\delta^{(n)}\} + (1 - \rho_{n+1})u^{(n-1)}, \tag{2.4}$$

where the *pseudo-residual vector* $\delta^{(n)}$ is defined by $\delta^{(n)} = Gu^{(n)} + k - u^{(n)}$. For Chebyshev acceleration $\gamma_1 = \gamma_2 = \cdots = \gamma$; the numbers $\gamma, \rho_1, \rho_2, \cdots$ can be determined in terms of $m(G)$ and $M(G)$ which are estimates of the smallest and largest eigenvalues of G, respectively. (See Hageman and Young [1981, p. 48].) For conjugate gradient acceleration the values of $\gamma_1, \gamma_2, \cdots$ and $\rho_1, \rho_2, \cdots$ can be computed in terms of certain inner products involving $\delta^{(n)}$ and $\delta^{(n-1)}$. (See Hageman and Young [1981, p. 147].)

The seven algorithms of ITPACK include the Jacobi, SSOR, and RS methods, each with Chebyshev and conjugate gradient acceleration, and the SOR method without acceleration. The algorithms include automatic, or adaptive, procedures for determining the necessary iteration parameters. They also include realistic procedures for deciding when $u^{(n+1)}$ is sufficiently close to the true solution of (2.1) so that the iteration process can be terminated. Detailed information on the algorithms can be found in the book by Hageman and Young [1981]; see also Grimes, Kincaid, and Young [1979]. The usage of the software package is described in the paper by Kincaid, Respess, Young and Grimes [1982].

3. Vectorization of ITPACK routines. In this section we discuss the programs currently in ITPACK from the standpoint of vectorization. We will discuss both the storage schemes and the algorithms themselves.

Let us now look first at the routines of ITPACK from the standpoint of vectorization. Evidently, the acceleration procedures defined by (2.4) are vectorizable provided that the basic iterative method is vectorizable.

The Jacobi method is clearly vectorizable and the RS method is also highly vectorizable. To see this, we rewrite (2.3) in the form

$$\begin{aligned} u_R^{(n+1)} &= -D_R^{-1}Hu_B^{(n)} + D_R^{-1}b_R, \\ u_B^{(n+1)} &= -D_B^{-1}Ku_R^{(n+1)} + D_B^{-1}b_B. \end{aligned} \tag{3.1}$$

Assuming D_R and D_B are of approximately the same size, the vector length will be approximately $N/2$.

The SOR method is not in general vectorizable. The basic step in the SOR method involves the solution of an auxiliary linear system with a lower triangular matrix. The solution of such a system can be carried out by a forward substitution procedure. This is efficient for a scalar machine but is clearly not efficient for a vector computer since in order to get the ith component of the solution of the auxiliary system one must have available the kth component for $k = 1, 2, \cdots, i-1$. There is, however, an important case where the SOR method can be vectorized. If the matrix A of (2.1) has *Property* A, (see Young [1971]), then one can permute the equations and relabel the unknowns so that one obtains a *red-black system* of the form (2.2). The SOR method is defined by

$$\begin{aligned} u_R^{(n+1)} &= \omega\{-D_R^{-1}Hu_B^{(n)} + D_R^{-1}b_R\} + (1-\omega)u_R^{(n)}, \\ u_B^{(n+1)} &= \omega\{-D_B^{-1}Ku_R^{(n+1)} + D_B^{-1}b_B\} + (1-\omega)u_B^{(n)}. \end{aligned}$$

Thus the computation of $u_R^{(n+1)}$ is vectorizable with vector length approximately $N/2$. Similarly, the computation of $u_B^{(n+1)}$ is vectorizable.

An important case where one obtains a matrix with Property A is when one is solving a five-point difference equation for a square mesh, derived from an elliptic boundary-value problem, for a two-dimensional region. Thus, if we consider the solution of Laplace's equation $u_{xx} + u_{yy} = 0$ on the square $0 \leqq x \leqq 1, 0 \leqq y \leqq 1$ with a

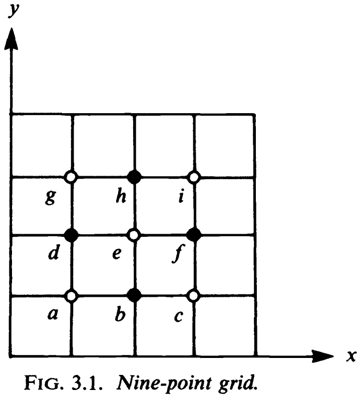


FIG. 3.1. *Nine-point grid.*

mesh size of $h = \frac{1}{4}$, we have the grid shown in Fig. 3.1. With the usual natural ordering of $\{a = 1,\ b = 2,\ c = 3,\ d = 4,\ e = 5,\ f = 6,\ g = 7,\ h = 8,\ i = 9\}$, the matrix of the system has Property A. However, if we designate the red points as a, c, e, g, i and the black points as b, d, f, h and label the points accordingly $\{a = 1,\ c = 2,\ e = 3,\ g = 4,\ i = 5,\ b = 6,\ d = 7,\ f = 8,\ h = 9\}$, then we get the following red-black matrix:

$$\begin{pmatrix} 4 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 4 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 4 & 0 & 0 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 & -1 & -1 \\ -1 & -1 & -1 & 0 & 0 & 4 & 0 & 0 & 0 \\ -1 & 0 & -1 & -1 & 0 & 0 & 4 & 0 & 0 \\ 0 & -1 & -1 & 0 & -1 & 0 & 0 & 4 & 0 \\ 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 4 \end{pmatrix}.$$

The SSOR method with the natural ordering is also not vectorizable. (We remark that one step of the SSOR method can be regarded as one iteration of the (forward) SOR method followed by one iteration of the backward SOR method.) If the matrix A has Property A, one could consider the corresponding red-black system. However, it is well-known, (see, for instance, Young [1971]), that the SSOR method is not effective when applied to a red-black system. This is in contrast to the SOR method which is no less effective for a red-black system than for the original system. Consequently, if one applies the SSOR method to a red-black system, one achieves vectorization but there is an increase in the number of iterations required for convergence which may not be off-set by the increase in speed.

For a five-point difference equation, it is possible to achieve some vectorization with the SSOR method without sacrificing the convergence rate. This can be done, as shown by Hayes [1977], by the use of the "ordering by diagonals". Here one lets $a = 1$, $b = 2$, $d = 3$, $c = 4$, $e = 5$, $g = 6, f = 7$, $h = 8$, $i = 9$. It can be shown that the ordering by diagonals is "equivalent" to the natural ordering for the SOR method. (See Young [1971].) Similarly in the backward sweep, we have equivalence. Thus, applying the SSOR method with the diagonal ordering gives the same convergence rate as with the natural ordering. On the other hand, there is now much greater vectorization than with the natural ordering since all values on a diagonal can be modified independently. Thus, the average vector length is the same as the average length of the diagonals.

The choice of the data structure for the coefficient matrix A has a significant impact on the vectorization of the routines of ITPACK. The scalar version of ITPACK uses the same data storage format as the Yale Sparse Matrix Package (YSMP) which uses three singly dimensioned arrays **A**, **JA**, and **IA**. (See Eisenstat et al. [1977].) In this data structure, **A** contains the nonzeros of the matrix stored by rows, **JA** contains the corresponding column numbers, and **IA** contains pointers into **A** and **JA** for the beginning locations of new rows. This data structure has great generality and can efficiently represent sparse matrices of random structure. However, this data scheme inhibits vectorization of basic operations such as computing a matrix-vector product because of the need to do indirect addressing and the short vector lengths involved. For example, in scalar ITPACK, a matrix-vector product for a system of size N was computed using N inner products applied to the condensed rows of the matrix, which are typically short.

To avoid this bottleneck on vector computers, it was decided to adopt the column oriented structure used in the ELLPACK software. (See Rice and Boisvert [1985].) In this data structure, two doubly dimensioned arrays, **COEF** and **JCOEF**, are used to store the matrix A. Each row of **COEF** contains the nonzero coefficients corresponding to a single equation and **JCOEF** contains the corresponding column numbers. Clearly, **COEF** and **JCOEF** must be dimensioned at least N by the maximum number of nonzeros per equation over all equations. This data structure is not as general as the YSMP structure since the possibility exists of storing a great number of zeros if one equation has many more nonzeros than the remaining equations. However, the matrix-vector product operation vectorizes with the use of gather/scatter instructions on the CYBER 205 and with the use of assembly-coded gather/scatter routines on the CRAY 1. Other operations, such as forward (back) substitutions using lower (upper) triangular matrices do not vectorize.

Another storage scheme we have investigated is storing the matrix A by diagonals. In this data structure, each column of **COEF** contains a diagonal of the matrix and **JCOEF** contains its corresponding distance from the main diagonal. With this structure, a matrix-vector product vectorizes without the use of gathering routines and operations such as forward (back) substitutions and factorizations vectorize to some extent. However, this storage format is the most rigid of the three since it can efficiently represent only matrices with a diagonal structure. Again, it is easily possible to store a great number of zeros if the matrix does not have a diagonal structure. This observation illustrates a trade-off we have frequently encountered—namely, increased vectorization often comes at the expense of more rigid storage requirements and more computations, many of which may be operations involving zeros.

4. Short-range vectorization of ITPACK. We now describe the short-range modifications of ITPACK and the numerical experiments which were run on the CYBER 205 and CRAY 1 vector computers.

We decided to investigate the vectorization of ITPACK by modifying the code in the following ways.

- Make a minimum number of changes to the current scalar ITPACK package to allow as much vectorization as possible. This generally meant re-rolling DO loops which had been unrolled for efficiency on scalar computers and using the available dot product routines on the vector computers. The tightly rolled loops were recognized by the compiler as vectorizable, thereby gaining improvement in the speed.
- Change the data structure of scalar ITPACK to one allowing greater vectorization of certain matrix-vector operations. This major rewriting effort was deemed necessary when it was discovered that the computations associated with the basic iterative method constituted a time-consuming bottleneck for vector computers. The resulting rewritten package will be referred to as "vector ITPACK".
- Make use of the vector syntax extensions to Fortran available on the CYBER 205 computer. It was hoped that the use of this syntax would improve the efficiency of the package.

These changes resulted in four different versions of the package which were used in the timing tests to be presented here.

Scalar ITPACK #1 is the standard scalar version as in Kincaid, Respess, Young, and Grimes [1982]. This version uses unrolled DO loops in basic vector operations for increased performance on scalar computers.

Scalar ITPACK #2 is the standard scalar version but with re-rolled DO loops and a few minor changes such as the use of the special dot product routines available on vector computers.

Vector ITPACK #1 is the package rewritten to use the ELLPACK data structure. This version uses standard Fortran primarily. The special dot product routines were used as well as the gathering and scattering instructions available on the CYBER 205 and the corresponding software routines on the CRAY 1 computer.

Vector ITPACK #2 is the CYBER 205 version of vector ITPACK #1 with heavy use of the vector syntax available in the CYBER Fortran.

The following test problem was used in this numerical experiment. A five-point finite-difference stencil was used to discretize the partial differential equation:

$$\begin{aligned} &u_{xx}(x, y) + 2u_{yy}(x, y) = 0, \qquad &&(x, y) \in S = (0, 1) \times (0, 1), \\ &u(x, y) = 1 + xy, &&(x, y) \in \text{boundary}(S). \end{aligned}$$

The mesh size chosen in the first experiment was $h = 1/64$, resulting in 3,969 unknowns. The stopping criterion was 5.0E-6. Both a natural ordering and a red-black ordering of the unknowns were used. While this is a rather simple problem, the timing results are felt to be representative of those from more complicated problems since the software does not take advantage of the constant coefficients in the partial differential equation.

Tables 1 and 2 give the iteration times for the CYBER 205 and CRAY computers, respectively. The time necessary to scale the system and permute the matrix (if red-black ordering was requested) is presented in Tables 3 and 4.

A number of observations can be made based on these results:

(i) A slight increase in speed resulted from removing scalar optimization tricks such as unrolling *DO* loops from the original scalar ITPACK. The increase was small since the bulk of the computations is the performance of the basic iterative method, which remains nonvectorizable in scalar ITPACK #2.

(ii) There was a considerable improvement in performance from scalar to vector versions of ITPACK. The ELLPACK data structure allowed the matrix-vector product operation to vectorize to a considerable extent, and this was the dominant computational kernel for many of the methods (the Jacobi methods for natural ordering and all methods for red-black ordering). The SOR and SSOR methods are predominantly

TABLE 1

Iteration time (*secs.*), CYBER 205 ($h = 1/64$).

Method	Iterations	Scalar ITPACK #1	Scalar ITPACK #2	Vector ITPACK #1	Vector ITPACK #2
		(Natural ordering)			
Jacobi CG	178	2.470	2.224	0.257	0.247
Jacobi SI	362	5.628	4.561	0.573	0.554
SOR	216	4.698	4.644	2.484	2.476
SSOR CG	34	2.128	1.790	0.843	0.839
SSOR SI	43	1.879	1.765	0.984	0.980
		(Red-black ordering)			
Jacobi CG	178	2.343	2.117	0.261	0.252
Jacobi SI	362	5.357	4.332	0.583	0.562
SOR	196	4.110	4.084	0.488	0.470
SSOR CG ($\omega = 1$)	70	3.785	3.123	0.209	0.194
SSOR SI	196	8.125	7.540	0.690	0.654
RS CG	90	1.456	1.358	0.116	0.108
RS SI	182	3.132	2.780	0.220	0.203

TABLE 2
Iteration time (*secs.*), CRAY 1 ($h = 1/64$).

Method	Iterations	Scalar ITPACK #1	Scalar ITPACK #2	Vector ITPACK #1
	(Natural ordering)			
Jacobi CG	178	2.577	2.564	0.716
Jacobi SI	362	5.418	5.252	1.415
SOR	216	4.399	4.356	3.112
SSOR CG	34	1.771	1.725	1.143
SSOR SI	43	1.660	1.651	1.343
	(Red-black ordering)			
Jacobi CG	178	2.198	2.176	0.710
Jacobi SI	362	4.632	4.448	1.402
SOR	196	3.729	3.691	0.743
SSOR CG ($\omega = 1$)	70	3.000	2.909	0.612
SSOR SI	196	6.810	6.793	1.632
RS CG	90	1.590	1.574	0.328
RS SI	182	3.292	3.231	0.656

TABLE 3
Total time-iteration time, CYBER 205 ($h = 1/64$).

Method	Scalar ITPACK #1	Scalar ITPACK #2	Vector ITPACK #1	Vector ITPACK #2
	(Natural ordering)			
Jacobi CG	.091	.082	.030	.030
Jacobi SI	.091	.081	.031	.030
SOR	.090	.082	.060	.060
SSOR CG	.091	.083	.060	.060
SSOR SI	.091	.082	.060	.060
	(Red-black ordering)			
Jacobi CG	.714	.704	.068	.066
Jacobi SI	.714	.704	.067	.067
SOR	.713	.703	.068	.067
SSOR CG ($\omega = 1$)	.715	.704	.068	.067
SSOR SI	.714	.704	.067	.066
RS CG	.721	.709	.069	.067
RS SI	.722	.710	.068	.068

recursive for natural ordering, and the improvement in performance was not as great in the vector versions of ITPACK.

(iii) There was a marginal improvement in speed in going from the standard Fortran version of vector ITPACK #1 to the CYBER 205 vector syntax version of vector ITPACK #2. For our applications, there were very few computations which could not be recognized as vectorizable by the CYBER 205 compiler when written in standard Fortran. The small savings in time resulted when calls to vector subroutines were replaced by in-line vector instructions.

(iv) Comparisons of methods based upon the number of iterations are misleading on vector computers. Methods which are slow to converge but are susceptible to

TABLE 4
Total time-iteration time, CRAY 1 ($h = 1/64$).

Method	Scalar ITPACK #1	Scalar ITPACK #2	Vector ITPACK #1
	(Natural ordering)		
Jacobi CG	.091	.090	.046
Jacobi SI	.091	.090	.045
SOR	.090	.090	.082
SSOR CG	.091	.090	.082
SSOR SI	.090	.090	.082
	(Red-black ordering)		
Jacobi CG	.717	.715	.106
Jacobi SI	.716	.715	.106
SOR	.716	.716	.106
SSOR CG ($\omega = 1$)	.717	.715	.105
SSOR SI	.717	.716	.106
RS CG	.725	.724	.108
RS SI	.726	.724	.108

vectorization can be more efficient than methods which have good convergence properties but involve recursive calculations. Hence, the JACOBI CG and JACOBI SI methods seem to be preferable to the recursive algorithms of the SOR, SSOR CG, and SSOR SI methods in the case of natural ordering. For red-black ordering, the RS methods seem to be the most efficient. It is interesting to note that SOR, SSOR CG, and SSOR SI methods vectorize with red-black ordering, thus performing better than with natural ordering. This remains true for the SSOR methods in spite of the greater number of iterations.

(v) The total time for each method is not significantly greater than the iteration time in the vector version, as can be seen from Tables 3 and 4. This result is due to the fact that the scaling and permuting operations on the matrix are also vectorizable with the ELLPACK column-oriented data structure.

Tables 5 and 6 give the time per iteration for each method using natural and red-black ordering for the CYBER 205 and CRAY computers, respectively.

Both the CYBER 205 and CRAY computers perform at about the same speed for this problem using the scalar versions of ITPACK. There seems to be less of an improvement in speed in going to the vectorized version of ITPACK for the CRAY computer than for the CYBER 205. For those methods which vectorize well (i.e., those methods whose basic iterative step is a matrix-vector multiply), the CYBER 205 achieves an order of magnitude improvement in speed for this problem size, while the CRAY 1 improves by a factor of three to five. This may be due to the CYBER 205's efficiency in processing long vectors and its hardware gathering and scattering instructions. For the data structure of vector ITPACK, gathering operations are a significant part of the matrix-vector multiply.

For both computers, there was not as significant an improvement in speed for the SOR and SSOR methods for the natural ordering of the unknowns. These methods require forward or back solutions through sparse triangular factors which are recursive with the ELLPACK data structure.

It was also decided to test vector ITPACK on the test problem listed above for various mesh sizes to determine the effect of the vector length on the performance.

TABLE 5
Time per iteration (secs.), CYBER 205 ($h = 1/64$).

Method	Scalar ITPACK #1	Scalar ITPACK #2	Vector ITPACK #1	Vector ITPACK #2
	(Natural ordering)			
Jacobi CG	.0139	.0125	.0014	.0014
Jacobi SI	.0155	.0126	.0016	.0015
SOR	.0218	.0215	.0115	.0115
SSOR CG	.0626	.0526	.0248	.0247
SSOR SI	.0437	.0410	.0229	.0228
	(Red-black ordering)			
Jacobi CG	.0132	.0119	.0015	.0014
Jacobi SI	.0148	.0120	.0016	.0016
SOR	.0210	.0208	.0025	.0024
SSOR CG ($\omega = 1$)	.0541	.0446	.0030	.0028
SSOR SI	.0415	.0385	.0035	.0033
RS CG	.0162	.0151	.0013	.0012
RS SI	.0172	.0153	.0012	.0011

TABLE 6
Time per iteration (secs.), CRAY 1 ($h = 1/64$).

Method	Scalar ITPACK #1	Scalar ITPACK #2	Vector ITPACK #1
	(Natural ordering)		
Jacobi CG	.0145	.0144	.0040
Jacobi SI	.0150	.0145	.0039
SOR	.0204	.0202	.0144
SSOR CG	.0521	.0507	.0336
SSOR SI	.0386	.0384	.0312
	(Red-black ordering)		
Jacobi CG	.0123	.0122	.0040
Jacobi SI	.0128	.0123	.0039
SOR	.0190	.0188	.0038
SSOR CG ($\omega = 1$)	.0429	.0416	.0087
SSOR SI	.0347	.0347	.0083
RS CG	.0177	.0175	.0036
RS SI	.0181	.0178	.0036

Both the CYBER 205 and CRAY computers are known to become increasingly efficient as the vector length grows since the start-up times for vector computations becomes increasingly insignificant relative to the stream time. The mesh sizes chosen were $h = 1/16$, $1/32$, $1/64$, $1/128$, and $1/256$, resulting in 225, 961, 3,969, 16,129, and 65,025 unknowns, respectively.

Table 7 gives the different number of iterations for each method and each mesh size. Tables 8 and 9 give the corresponding iteration times for the CYBER 205 and CRAY 1 computers, respectively. Tables 10 and 11 give the iteration time per node per iteration. The largest problem could not be run on the particular CRAY 1 being used because of limited available memory.

TABLE 7
Iterations

Method	$h = 1/16$	1/32	1/64	1/128	1/256
		(Natural ordering)			
Jacobi CG	49	94	178	330	629
Jacobi SI	84	179	362	772	1372
SOR	50	104	216	422	872
SSOR CG	16	22	34	51	73
SSOR SI	19	29	43	61	88
		(Red-black ordering)			
Jacobi CG	49	94	178	330	629
Jacobi SI	84	179	362	772	1372
SOR	52	101	196	396	839
SSOR CG ($\omega = 1$)	20	37	70	120	223
SSOR SI	51	107	196	373	752
RS CG	25	48	90	167	321
RS SI	42	88	182	375	704

TABLE 8
Iteration time (*secs.*), CYBER 205.

Method	$h = 1/16$	1/32	1/64	1/128	1/256
		(Natural ordering)			
Jacobi CG	.010	.040	.247	1.792	14.121
Jacobi SI	.018	.091	.554	4.180	29.919
SOR	.036	.296	2.476	19.545	165.841
SSOR CG	.028	.136	.839	5.022	28.564
SSOR SI	.029	.166	.980	5.663	32.736
		(Red-black ordering)			
Jacobi CG	.010	.040	.252	1.824	14.395
Jacobi SI	.018	.090	.562	4.259	29.211
SOR	.011	.066	.470	3.745	32.199
SSOR CG ($\omega = 1$)	.007	.031	.194	1.293	9.893
SSOR SI	.020	.112	.654	4.372	35.688
RS CG	.006	.019	.108	.749	5.949
RS SI	.008	.032	.203	1.540	11.804

Tables 10 and 11 give the approximate time spent per node on each iteration. On scalar computers, it would be expected that this quantity would be independent of the problem size. However, as can be seen above, this quantity decreases as the problem size grows for those methods which vectorize well. Hence, the JACOBI CG and JACOBI SI methods for natural ordering and all methods for red-black ordering improve their efficiency as the problem size grows, while methods which vectorize poorly (the SOR and SSOR methods for natural ordering) show little improvement in efficiency. The figures also indicate that for this experiment the CYBER 205 had a greater improvement in efficiency for long vector computations than did the CRAY 1.

The gather/scatter operations on the CRAY 1 are done in software, whereas the CYBER 205 has special hardware instructions for them. In fact, the CRAY 1 has carefully written assembly code for gathering and scattering which runs faster than the corresponding Fortran code but true hardware gather/scatter instructions such as

TABLE 9
Iteration time (secs.), CRAY 1.

Method	$h = 1/16$	1/32	1/64	1/128
	(Natural ordering)			
Jacobi CG	.015	.098	.716	5.296
Jacobi SI	.024	.183	1.415	11.899
SOR	.044	.369	3.112	24.576
SSOR CG	.034	.184	1.143	6.849
SSOR SI	.037	.224	1.343	7.724
	(Red-black ordering)			
Jacobi CG	.014	.097	.710	5.259
Jacobi SI	.025	.182	1.402	11.812
SOR	.015	.100	.743	5.982
SSOR CG ($\omega = 1$)	.013	.085	.612	4.163
SSOR SI	.032	.237	1.632	12.076
RS CG	.008	.047	.328	2.408
RS SI	.013	.086	.656	5.305

on some CRAY X-MP's should run substantially faster still. This many explain why vector ITPACK enjoys a much greater speedup on the CYBER 205 than on the CRAY 1.

Many iterative methods in ITPACK were susceptible to vectorization, but a major rewriting of the package would be necessary for a vector computer to "notice" all vectorization possibilities. It is generally true that any code must be tailored to a particular computer in order to achieve optimum efficiency, but the potential reward for doing so on a vector computer is often an improvement of an order of magnitude increase in speed. Our goal in carrying out short-range modifications on ITPACK was to exploit the vector processing capabilities of vector computers without destroying the general purpose nature of the package. It is hoped that the choice of the ELLPACK data structure strikes a useful compromise between the demand for speed and the demand for flexibility.

5. Long-range modifications. In this section we give a brief description of some of our work on long-range modifications to ITPACK. These modifications will involve the adaptation of some current algorithms and the use of new algorithms.

As indicated in § 4, if one uses the natural ordering, the SSOR method is not vectorizable though it is often rapidly convergent. On the other hand, if one uses the red-black ordering, it is vectorizable and converges in a greater number of iterations which may be off-set by the increased speed. This behavior which is, as stated in § 3, caused by the fact that the solution of a linear system with a sparse triangular matrix is often not vectorizable, is typical of other approximate factorization methods, such as the Incomplete Cholesky method of Meijerink and Van der Vorst [1977]. For such methods one represents the matrix A of (2.1) in the form LU where L and U are sparse lower and sparse upper triangular matrices, respectively. The repeated solution of linear systems involving the matrices L and U are required.

In some cases an improvement in the vectorization can be made by reordering the rows and corresponding columns of A, and hence of L and U. For the SSOR method for a linear system corresponding to a 5-point finite difference equation one can, as stated in § 3, use ordering by diagonals to achieve some vectorization. We are planning to implement this in the near future not only for the SSOR method but for

TABLE 10
Time per iteration per node (microseconds), CYBER 205.

Method	$h = 1/16$	1/32	1/64	1/128	1/256
	(Natural ordering)				
Jacobi CG	.907	.443	.350	.337	.345
Jacobi SI	.952	.529	.386	.336	.335
SOR	3.200	2.962	2.888	2.872	2.925
SSOR CG	7.778	6.433	6.217	6.105	6.017
SSOR SI	6.784	5.956	5.742	5.756	5.721
	(Red-black ordering)				
Jacobi CG	.907	.443	.357	.343	.352
Jacobi SI	.952	.523	.391	.342	.327
SOR	.940	.680	.604	.586	.590
SSOR CG ($\omega = 1$)	1.556	.872	.698	.668	.682
SSOR SI	1.743	1.089	.841	.727	.730
RS CG	1.067	.412	.302	.278	.285
RS SI	.847	.378	.281	.255	.258

TABLE 11
Time per iteration per node (microseconds), CRAY 1.

Method	$h = 1/16$	1/32	1/64	1/128
	(Natural ordering)			
Jacobi CG	1.361	1.085	1.013	0.995
Jacobi SI	1.270	1.064	0.985	0.956
SOR	3.911	3.692	3.630	3.611
SSOR CG	9.444	8.703	8.470	8.326
SSOR SI	8.655	8.038	7.869	7.851
	(Red-black ordering)			
Jacobi CG	1.270	1.074	1.005	.988
Jacobi SI	1.323	1.058	.976	.949
SOR	1.282	1.030	.955	.937
SSOR CG ($\omega = 1$)	2.889	2.391	2.203	2.151
SSOR SI	2.789	2.305	2.098	2.007
RS CG	1.422	1.019	.918	.894
RS SI	1.376	1.017	.908	.877

other approximate factorization methods as well. Another approach we plan to undertake is based on work of Kershaw [1982] on the solution of block tridiagonal systems.

As an alternative to the use of an approximate factorization of A wherein A is represented as the product of sparse triangular factors, we are investigating the use of approximate inverses wherein A^{-1} is represented by a sparse matrix. Sparse approximate inverses have been developed by Dubois, Greenbaum and Rodrigue [1979] using a Neumann series and by Johnson, Michelli and Paul [1983] using more general polynomials.

The approach we have taken involves the construction of a sparse approximate inverse H based on the use of the Gaussian elimination method. To compute the columns $k^{(1)}, k^{(2)}, \cdots, k^{(N)}$ of the exact inverse A^{-1}, one solves a set of linear systems of the form

$$Ak^{(i)} = e^{(i)},$$

where $i = 1, 2, \cdots, N$, and where $e^{(i)}$ is the ith unit vector. Instead of this we determine an approximate inverse $\hat{H}$ such that $\hat{h}_{i,j} = 0$ if $(i, j) \notin S$. Here S is a sparsity set, i.e. a subset of the pair (i, j) such that the $i \leqq i, j \leqq N$. For each $i = 1, 2, \cdots, N$, we require that

$$A\hat{h}_j^{(i)} = e_j^{(i)}$$

provided that $(i, j) \in S$. Actually we can write

$$A^{(i)}\hat{h}_j^{(i)} = e_j^{(i)},$$

where $A^{(i)}$ is obtained from A by deleting certain rows and columns of A. In some applications the $A^{(i)}$ are small matrices and the columns $\hat{h}^{(i)}$ can be computed quickly in vector mode. The basic iterative method corresponding to a sparse approximate inverse is fully vectorizable.

As an example the matrix

$$A = \begin{pmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix}$$

with the sparsity set

$$S = \{1, 1), (1, 2)(1, 3)(2, 1), (2, 2), (2, 4), (3, 1), (3, 3), (3, 4), (4, 2), (4, 3), (4, 4)\}$$

has the incomplete inverse

$$\hat{H} = \begin{pmatrix} 2/7 & 1/14 & 1/14 & 0 \\ 1/14 & 2/7 & 0 & 1/14 \\ 1/14 & 0 & 2/7 & 1/14 \\ 0 & 1/14 & 1/14 & 2/7 \end{pmatrix}.$$

We remark that the determination of $\hat{H}$ can be carried out explicitly for the case of a five-point difference equation over a rectangular mesh; for details see Kincaid, Oppe and Young [1984]. Numerical experiments are currently underway to determine the effectiveness of the method for various choices of S.

6. Acknowledgments. We sincerely thank David Gay for carrying out the numerical tests on the CRAY 1 at Bell Laboratories. The CYBER 205 results were obtained by using the computer at Colorado State University.

REFERENCES

P. F. Dubois, A. Greenbaum and G. H. Rodrigue (1979), *Approximating the inverse of a matrix for use in iterative algorithms on vector processors*, Computing, 22, pp. 257-268.

S. C. Eisenstat, M. C. Gursky, H. H. Schultz and A. H. Sherman (1977), *Yale Sparse Matrix Package* II, *The Nonsymmetric Codes*, Report 114, Dept. Computer Science, Yale Univ., New Haven, CT.

S. Eisenstat, A. George, R. Grimes, D. Kincaid and A. Sherman (1979), *Some Comparisons of Software Packages for Large Sparse Linear Systems*, in Advances in Computer Methods for Partial Differential Equations III, R. Vichnevetsky and R. Stepleman, eds, IMACS Publ., Dept. Computer Science, Rutgers Univ., New Brunswick, NJ, pp. 98-106.

R. G. Grimes, D. R. Kincaid and D. M. Young (1979), ITPACK 2.0, *User's Guide*, Report CNA-150, Center for Numerical Analysis, Univ. Texas, Austin.

L. A. Hageman and D. M. Young (1981), *Applied Iterative Methods*, Academic Press, New York.

L. J. HAYES (1977), *Comparative analysis of iterative techniques for solving Laplace's equation on the unit square on a parallel processor*, M.A. thesis, Dept Mathematics, Univ. Texas, Austin.

O. G. JOHNSON, C. A. MICHELLI AND G. PAUL (1983), *Polynomial preconditioners for conjugate gradient calculations*, SIAM J. Numer. Anal., 20, pp. 362-376.

D. KERSHAW (1982), *Solution of single tridiagonal linear systems and vectorization of the* ICCG *algorithm on the Cray*-1, Parallel Computations, Garry Rodrigue, ed., Academic Press, New York, pp. 85-99.

D. R. KINCAID AND D. M. YOUNG (1979), *Survey of iterative methods*, in Encyclopedia of Computer Sciences and Technology 13, J. Belzer, A. Holzman and A. Kent, eds., Marcel Dekker, New York, pp. 354-391.

D. R. KINCAID, J. R. RESPESS, D. M. YOUNG AND G. R. GRIMES (1982), ITPACK 2C: *a* FORTRAN *package for solving large sparse linear systems by adaptive accelerated iterative methods*, ACM Trans. Math. Software, 8, pp. 302-322.

D. R. KINCAID, T. C. OPPE AND D. M. YOUNG (1982), *Adapting* ITPACK *routines for use on a vector computer*, Report CNA-177, Center for Numerical Analysis, Univ. Texas, Austin.

D. R. KINCAID AND T. C. OPPE (1983), ITPACK *on supercomputers*, in Numerical Methods, A. Dold and B. Echmann, eds., Springer-Verlag, Lecture Notes in Mathematics 1005, New York, pp. 151-161.

D. R. KINCAID AND D. M. YOUNG (1983), *The* ITPACK *project: Past, present, and future*, Report CNA-180, Center for Numerical Analysis, Univ. Texas, Austin.

D. R. KINCAID, T. C. OPPE AND D. M. YOUNG (1984), *Vector computations for sparse linear systems*, Report CNA-189, Center for Numerical Analysis, Univ. Texas, Austin.

D. R. KINCAID, T. C. OPPE, J. R. RESPESS AND D. M. YOUNG (1984), ITPACKV 2C, *User's Guide*, Report CNA-191, Center for Numerical Analysis, Univ. Texas, Austin.

J. A. MEIJERINK AND H. A. VAN DER VORST (1977), *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31, pp. 148-162.

J. R. RICE AND R. F. BOISVERT (1985), *Solving Elliptic Problems Using* ELLPACK, Springer-Verlag, New York.

D. M. YOUNG (1971), *Iterative Solution of Large Linear Systems*, Academic Press, New York.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

013

A GENERALIZED PARITY FUNCTION AND ITS USE IN THE CONSTRUCTION OF PERFECT CODES*

M. MOLLARD†

Abstract. We define a new generalized parity function, and use it to obtain a product construction of single-error-correcting codes (binary or not).

Key words. combinatorics, error-correcting codes

1. Introduction. Let V_n be the vector space of dimension n over the finite field $GF(q)$. A code of length n is a subset C of V_n. If C is a subspace of V_n the code is called linear. A single-error-correcting code (or distance 3 code) is a code C having the property

$$\forall x \in C, \quad \forall y \in C, \quad d(x, y) \geqq 3, \quad \text{or } x = y,$$

where $d(x, y)$ is the Hamming distance between x and y. This is the only kind of code considered here. The Hamming bound states that

$$|C| \leqq \frac{q^n}{n(q-1)+1}. \tag{1}$$

C is called perfect if (1) is an equality. Perfect codes of length n exist if and only if for some m

$$n = \frac{q^m - 1}{q - 1}.$$

The earliest examples of perfect codes are linear ones (Hamming [1] for $q = 2$, Shapiro and Slotnick [2] for the general case). Nonlinear perfect codes have been constructed by Vasiliev [3] ($q = 2$) and Schönheim [4]. Other nonlinear codes are also known.

In this paper we present a product construction for combining codes, based on the use of a "generalized parity function". Starting with binary single-error-correcting codes of lengths n, m this construction gives us a single-error-correcting code of length $nm + n + m$. But our purpose here is to construct perfect codes, so we will only apply it to such codes.

Phelps [5] has also introduced a product construction which generalizes his "combinatorial construction of perfect codes" [6]. However, when compared to Vasiliev's construction, Phelps' appears as being of a very different nature: we think that, more than likely, it cannot produce the Vasiliev codes, while their generalization is precisely our aim.

2. Product construction in the binary case. Let n_1, n_2 be integers.

DEFINITION. The generalized parity function $(P_1(x), P_2(x))$ from $V_{n_1 n_2}$ to $V_{n_1+n_2}$ is defined by

$$P_1(x) = \left(\sum_{j=1}^{n_2} x_{ij} \right), \qquad i \in \{1, \cdots, n_1\},$$

$$P_2(x) = \left(\sum_{i=1}^{n_1} x_{ij} \right), \qquad j \in \{1, \cdots, n_2\},$$

* Received by the editors December 15, 1983 and in revised form November 26, 1984.

† Laboratoire LSD (IMAG), 38402, St. Martin d'Hères Cédex, France.

where the components of x, an element of $V_{n_1n_2}$, are $(x_{11}, \cdots, x_{1n_2}, x_{21}, \cdots, x_{n_1n_2})$, i.e. the coordinate positions of x are arranged in lexicographic order over $\{1<\cdots<n_1\}\times\{1<\cdots<n_2\}$.

Let C and C' be perfect codes of lengths n_1, n_2, and let f be a vector function from C to V_{n_2}. Now define F as

$$F=\{(x, c+P_1(x), c'+P_2(x)+f(c))\},$$

where $x\in V_{n_1n_2}$, $c\in C$, and $c'\in C'$.

THEOREM 1. *F is a single-error-correcting perfect code of length $n=n_1n_2+n_1+n_2$.*

Remark. Two important particular cases are:

1) f is a constant value function,

2) $n_2=1$, then $P_1(x)=x$, $P_2(x)=P(x)$ the classical parity function; this is Vasiliev's construction.

Proof of Theorem 1. First notice that for some a and b

$$n_1=2^a-1, \qquad n_2=2^b-1,$$

and so

$$n=n_1n_2+n_1+n_2=2^{a+b}-1.$$

The number of vectors in F is

$$|F|=2^{n_1n_2}\frac{2^{n_1}}{n_1+1}\frac{2^{n_2}}{n_2+1}=\frac{2^n}{n+1}.$$

Therefore, if F is single-error-correcting, it must be perfect. Let a and $\bar{a}$ be two different vectors of F. We have to show that $d(a, \bar{a})\geqq 3$. For some x, $\bar{x}$, c, $\bar{c}$, c', $\bar{c}'$ we can write

$$a=(x, c+P_1(x), c'+P_2(x)+f(c)),$$
$$\bar{a}=(\bar{x}, \bar{c}+P_1(\bar{x}), \bar{c}'+P_2(\bar{x})+f(\bar{c})).$$

a) If $x=\bar{x}$, then $P_1(x)=P_1(\bar{x})$, $P_2(x)=P_2(\bar{x})$ and $d(a, \bar{a})=d(c, \bar{c})+d(c', \bar{c}')\geqq 3$.

b) If $d(x, \bar{x})=1$, then $d(P_1(x), P_1(\bar{x}))=d(P_2(x), P_2(\bar{x}))=1$. If $c\neq\bar{c}$, then $d(c+P_1(x), \bar{c}+P_1(\bar{x}))\geqq 2$ and $d(a, \bar{a})\geqq 3$. If $c=\bar{c}$, then $d(\bar{c}'+P_2(\bar{x})+f(\bar{c}), c'+P_2(x)+f(c))\geqq 1$ and again $d(a, \bar{a})\geqq 3$.

c) If $d(x, \bar{x})=2$, then $d(P_1(x), P_1(\bar{x}))$ and $d(P_2(x), P_2(\bar{x}))$ are 0 or 2 but both cannot be zero at the same time. Therefore, the equalities

1) $c+P_1(x)=\bar{c}+P_1(\bar{x})$,

2) $c'+P_2(x)+f(c)=\bar{c}'+P_2(\bar{x})+f(\bar{c})$,

are not compatible and $d(a, \bar{a})\geqq 3$.

d) The trivial case $d(x, \bar{x})\geqq 3$ ends the proof.

Let $C_1, C_2, \cdots, C_p$ be perfect codes of lengths $n_1, n_2, \cdots, n_p$ and let

$$m=\prod_{i=1}^{p}(n_i+1)-\sum_{i=1}^{p}n_i-1.$$

We can define a family of functions R_k ($k\in\{1, 2, \cdots, p\}$) from V_m to V_{n_k}, playing the part of a parity function, and use it to build perfect codes of length $\prod_{i=1}^{p}(n_i+1)-1$. The interested reader can find this generalization of the above construction in [8].

3. General construction. Finally, we are going to state a generalization to perfect codes over finite fields, and so generalize Schönheim's construction. A different generalization, over arbitrary alphabets, recently has been developed by Phelps [7].

Let q be a power of a prime and let a_i $(i=1, 2, \cdots, q-1)$ be the nonzero elements of $GF(q)$ in a fixed order. For two integers n_1, n_2 arrange the coordinate positions of x, a word of $V_{(q-1)n_1n_2}$ in the lexicographic order over

$$\{1, \cdots, q-1\} \times \{1, \cdots, n_1\} \times \{1, \cdots, n_2\}.$$

$P_1(x)$ will be the function from $V_{(q-1)n_1n_2}$ to V_{n_1} defined by

$$P_1(x) = (y_1, \cdots, y_j, \cdots, y_{n_1}),$$

where

$$y_j = \sum_{i=1}^{q-1} \sum_{k=1}^{n_2} x_{ijk}.$$

$P_2(x)$ will be from $V_{(q-1)n_1n_2}$ to V_{n_2} defined by

$$P_2(x) = (y_1, \cdots, y_k, \cdots, y_{n_2}),$$

where

$$y_k = \sum_{i=1}^{q-1} a_i \sum_{j=1}^{n_1} x_{ijk}.$$

THEOREM 2. *Let C and C' be two perfect codes of lengths n_1, n_2 and let f be a function from C to V_{n_2}. F is a perfect code over $V_{(q-1)n_1n_2+n_1+n_2}$ where*

$$F = \{(x, c+P_1(x), c'+P_2(x)+f(c)), \qquad x \in V_{(q-1)n_1n_2}, \quad c \in C, \quad c' \in C'.$$

A proof of this theorem can be found in [8]. We only remark here that for $n_2 = 1$ we obtain the Schönheim nonlinear perfect codes.

REFERENCES

[1] R. W. HAMMING, *Error detecting and error correcting codes*, Bell Syst. Tech. J., 29 (1950), pp. 147-160.

[2] H. S. SHAPIRO AND D. L. SLOTNICK, *On the mathematical theory of error correcting codes*, IBM J. Res. Develop., 3 (1959), pp. 25-37.

[3] L. VASILIEV JR., *On nongroup close-packed codes*, 8 (1962), pp. 337-339.

[4] J. SCHÖNHEIM, *On linear and nonlinear single-error-correcting q-nary perfect codes*, INFORM. and Control, 12 (1968), pp. 23-26.

[5] K. T. PHELPS, *A general product construction for error correcting codes*, this Journal, 5 (1984), pp. 224-228.

[6] ———, *A combinatorial construction of perfect codes*, this Journal, 4 (1983), pp. 398-403.

[7] ———, *A product construction for perfect codes over arbitrary alphabets*, IEEE Trans. Inform. Theory, to appear.

[8] M. MOLLARD, *Une généralisation de la fonction parité, application à la construction de codes parfaits*, R.R. no 395, Laboratorire de Mathématiques Appliquées de Grenoble, Grenoble, France.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

014

THE CHARACTERS OF THE INFINITE SYMMETRIC GROUP AND PROBABILITY PROPERTIES OF THE ROBINSON–SCHENSTED–KNUTH ALGORITHM*

SERGEI V. KEROV† AND ANATOL M. VERSHIK†

In memory of V. A. Rohlin

Abstract. Connections between the Robinson-Schensted-Knuth algorithm, random infinite Young tableaux, and central indecomposable measures are investigated. A generalization of the RSK algorithm leads to a combinatorial interpretation of extended Schur functions. Applications are given to Ulam's problem on longest increasing subsequences and to a law of large numbers for representations. An analogous theory for other graphs is discussed.

Key words. Young tableaux, random infinite Young tableaux, extended Schur functions, RSK-algorithm, law of large numbers for representations of symmetric groups

1. Introduction. In the past several years the remarkable Robinson-Schensted-Knuth (RSK) algorithm has found numerous applications (see e.g. [2]). In this article we apply the algorithm (more exactly its "right half") to infinite sequences of letters of a certain linearly ordered alphabet corresponding to infinite Young tableaux. Then considering the elements of the sequences as independent random letters (with the same distribution on the alphabet) we get infinite random Young tableaux. The main result (Theorem 2) asserts that the measures on Young tableaux arising in this way (i.e. the images of the product-measures) are the central (see § 6) indecomposable measures.

If we use a generalization of the RSK algorithm in which we divide the alphabet into "positive and negative" letters (as it will be done in § 2) *the list of the indecomposable central measures will be completely exhausted by the images of the product-measures under the* RSK *algorithm.* This generalized RSK algorithm was independently found by Berele and Regev [1] and further considered in [10].

On the other hand every central measure on the space of infinite Young tableaux defines a character of the factor-representation of finite type of the group $\mathfrak{S}_\infty$ of all finite permutations of positive integers (see § 6). Thus we get some new information about the finite characters and the corresponding representations of $\mathfrak{S}_\infty$ and about symmetric groups $\mathfrak{S}_n$ with large degree n. In addition, Theorems 1 and 2 have many probabilistic and combinatorial corollaries (see §§ 6–8).

In the paper [17] (see also [8]) an important special case had been considered: the image of the product-measure with purely continuous multiplier is the Plancherel measure on infinite Young tableaux which corresponds to the regular representation character of $\mathfrak{S}_\infty$. Its statistical properties (including the limit shape of typical random Young tableaux) were studied in [17] and [8]; and a number of combinatorial corollaries, including the complete solution of Ulam's problem on the expected length of the longest increasing subsequence of a random permutation, were given in [17] and will be mentioned here in § 7.

The generalization of the RSK algorithm mentioned above is a combination of two RSK variants described in [7]. This modified algorithm, defined in § 2, is very important by itself even for finite n; it conserves the usual properties of the RSK

* Received by the editors November 26, 1984.

† Mathematics Department, Leningrad State University, Leningrad 198 904, USSR.

algorithm and connects combinatorial aspects of Young tableaux with the theory of symmetric functions and representations of $\mathfrak{S}_n$. As an example we shall give in § 5 an interesting probabilistic interpretation of the Schur functions.

The corollaries of the main theorem break up into various parts: *new laws of large numbers* for nonlinear functionals on the sequences of independent random variables (§ 7), asymptotic primarity of induced characters of $\mathfrak{S}_n$, "*the law of large numbers*" *for expansions of the induced representations on primary components* (§ 8). Some combinatorial corollaries are contained in § 7. Theorem 2 explains the reason why the set of indecomposable characters has the structure of a simplex and reveals the meaning of the parameters of those characters in Thoma's formula [14]—viz, *the frequencies of the rows and columns* (cf. [6]).

The probabilistic properties of the projection which we call "youngization" for the infinite case need further study; we establish here only that youngization is a homomorphism of spaces with invariant measure. Both the structure of the partition on the preimage of points and the description of measures invariant under Knuth's transformations (see § 3) are very interesting unsolved questions.

It is useful to look at a Young tableau as a path in Young's graph. From this viewpoint there can be seen an important generalization of the RSK algorithm in a completely different direction. In § 9 we shall give new examples of graded graphs and projections from the space of sequences into the space of paths in these graphs for which our main theorem is justified, viz., the list of the indecomposable central measures in the latter space is exhausted by the image of the product measures under this projection. Kingman's results on so-called partition structures [6] and the graph of finite ideals in a binary tree which Stanley [12] has studied, are included into our considerations. To find other interesting examples and a corresponding RSK algorithm is an open problem. For the connections with the theory of representations of $\mathfrak{S}_\infty$ see [14], [17], [16].

2. Bitabulation. Denote by $\mathbb{N}$ the set of positive integers; the Young diagram $\lambda \vdash n$ is an order ideal with n elements (cells) in $\mathbb{N} \times \mathbb{N}$ provided with the usual partial order. Let L be a linearly ordered alphabet with the partition $L = L_e \cup L_0$ (for our aims it is enough to put $L \subset \mathbb{R}$, $L_e = L \cap \mathbb{R}^+$). For $x, y \in L$ we shall write $x \nearrow y$ if $x < y$ or $x = y \in L_e$ and $x \searrow y$ in other cases. We shall call a word $w = x_1 x_2 \cdots x_n \in L^n$ *increasing* if $x_1 \nearrow x_2 \nearrow \cdots \nearrow x_n$ and *decreasing* if $x_1 \searrow x_2 \searrow \cdots \searrow x_n$.

Let $t \colon \lambda \to L$ be a map, corresponding the letters of L to the cells of a diagram λ. We shall call t an L-tableau, if the letters are increasing along the rows and decreasing along the columns of λ in the above sense (see the example below).

We define the algorithm INS to consist of the successive substitutions of the letters $x \in L$ of a word w into the rows of the tableau t; by definition the insertion of a single letter x coincides with INSERT from [7] if $x \in L_e$ and with INSERT* if $x \in L_0$. The inverse operation DEL combines DELETE and DELETE* from [7] in the same way.

Using this algorithm INS we associate with a word $w = x_1 \cdots x_n$ in the alphabet L a bitableau (R_n, S_n) defined as follows. The first tableau R_n is the result of applying INS to the letter x_n and the tableau R_{n-1}, the result being an L-tableau filled with the letters of L; the second one S_n is a usual Young tableau with integers $1, 2, \cdots$ as elements in its cells which is obtained from S_{n-1} by placing the integer n into the new cell of the diagram of R_n (so that R_n and S_n have the same shape). We shall call R_n the *sorting* and S_n the *accompanying* tableau of w. We put $R_0 = S_0 = \varnothing$ (the tableau with empty diagram); thus by induction we have defined R_n, S_n completely. Let $\Delta(w)$ be the common diagram (shape) of R_n, S_n.

Example. If $w=-2, 2, -2, 1, 3, -1, 3$ then

$$R_7(w)=\begin{array}{llll} 2 & & & \\ -2 & 1 & & \\ -2 & -1 & 3 & 3 \end{array} \qquad S_7(w)=\begin{array}{llll} 6 & & & \\ 3 & 4 & & \\ 1 & 2 & 5 & 7 \end{array}$$

All the results of [7] are preserved (after suitable changing) for our modification of the RSK algorithm. In particular, the operation of bitabulation defines a *bijection* between the spaces of n-words $w\in L^n$ and the set of pairs (R, S) where R is an L-tableau and S a Young tableau with the same diagram. The transformation which corresponds to the word w the tableau $S(w)$ we shall call *youngization.*

3. Generalized Schensted's theorem. Let w be a word in the alphabet L and λ an arbitrary Young diagram. We denote the maximal cardinality of the union of k increasing (resp. decreasing) subsequences of the word w by $r_k(w)(c_k(w))$, and let $\tilde{r}_k(\lambda)(\tilde{c}_k(\lambda))$ be the sum of the lengths of the first k rows (resp. columns) of the Young diagram λ.

PROPOSITION 1. *For any $w\in L^n$ and $k=1, 2, \cdots, n$, we have*

$$r_k(w)=\tilde{r}_k(\Delta(w)), \qquad c_k(w)=\tilde{c}_k(\Delta(w)),$$

where $\Delta(w)$ is the common diagram of $R(w)$ and $S(w)$.

It is not difficult to get the proof using Knuth's equivalence [7, Thm. 6], which has the following form in our case:

$$zxy\equiv xzy \quad \text{if } x<y<z \text{ or } x<y=z\in L_0 \text{ or } L_e\ni x=y<z, \tag{1}$$

$$yxz\equiv yzx \quad \text{if } x<y<z \text{ or } x<y=z\in L_e \text{ or } L_0\ni x=y<z. \tag{2}$$

Following [7] it is easy to check that $R(w')=R(w)$ if we can get the word w' from the word w by a chain of modifications of the form (1) or (2). The generalization of Proposition 1 to arbitrary finite partially ordered sets was obtained by Green and Kleitman [4] and independently by S. Fomin [3].

4. Youngization. We shall denote the set of Young tableaux with n cells by T_n (by definition a Young tableau is an increasing chain of diagrams $\phi=\lambda_0\subset\lambda_1\subset\cdots\subset\lambda_n$, $|\lambda_k|=k$), and by T the space of infinite Young tableaux with the topology of the projective limit: $T=\varprojlim T_n$.

Let L^∞ be the space of infinite sequences of letters of the alphabet L. For $w\in L^\infty$ we denote by w_n the initial piece of length n of the word w, and we let $S(w)\in T$ denote the infinite Young tableau $(\phi, \Delta_1(w_1), \cdots, \Delta_n(w_n), \cdots)$, defined as the limit of the accompanying finite tableaux $S_n(w_n)$. Note that the limit of the sorting tableaux does not exist in general. We can also consider infinite tableaux as infinite paths in the graded graph of finite Young diagrams (see § 9). The transformation S: $L^\infty\to T$ which we have constructed will also be called youngization; it is continuous and surjective.

5. The symmetric functions as probabilities. We shall say that a probability measure d on L has type (α, β, γ) where $\alpha=(\alpha_1, \alpha_2, \cdots)$, $\beta=(\beta_1, \beta_2, \cdots)$, α_k, β_k, $\gamma\in\mathbb{R}^+$, $\alpha_1\geqq\alpha_2\geqq\cdots$, $\beta_1\geqq\beta_2\geqq\cdots$, if $\alpha_1, \alpha_2, \cdots$ are the values of its atoms on the points of L_e and $\beta_1, \beta_2, \cdots$ are the values of atoms on the points of L_0, and finally if γ is the value of the measure on its continuous part, so that $\Sigma\alpha_k+\Sigma\beta_k+\gamma=1$.

Denote by $h_n(\alpha, \beta, \gamma)$ the probability that the random word $w=x_1\cdots x_n$ with independent letters with the common distribution d of type (α, β, γ) is increasing in the sense of § 2.

PROPOSITION 2. *The generating series for the function h_n has the following form*:

$$1+\sum_{n=1}^{\infty} h_n(\alpha, \beta, \gamma)z^n = e^{\gamma z} \prod_{i\geqq 1} \frac{1+\beta_i z}{1-\alpha_i z}.$$

(*For $\gamma=\beta_1=\beta_2=\cdots=0$ we get the complete homogeneous symmetric functions*

$$h_n = \sum_{i_1\leqq i_2\leqq\cdots\leqq i_n} \alpha_{i_1}\alpha_{i_2}\cdots\alpha_{i_n} \quad .)$$

As every symmetric function f has a unique representation as a polynomial in the arguments h_n, we can define its extension as the result of substitution of the function $h_n(\alpha, \beta, \gamma)$ at the place of h_n in this polynomial, for $n=1, 2, \cdots$. For instance the extended power sum symmetric functions are the following:

$$s_1(\alpha, \beta, \gamma) = \sum_{k=1}^{\infty} \alpha_k + \sum_{k=1}^{\infty} \beta_k + \gamma,$$

$$s_n(\alpha, \beta, \gamma) = \sum_{k=1}^{\infty} \alpha_k^n + (-1)^{n+1} \sum_{k=1}^{\infty} \beta_k^n, \qquad n \geqq 2.$$

We shall denote by e_λ the extended Schur function corresponding to the Young diagram λ. When $\gamma=0$, e_λ coincides with the hook-Schur functions HS_λ of [1] and [10], and the super-Schur functions $s_\lambda(\alpha/\beta)$ of [13]. The probabilistic meaning of extended Schur functions is given by the next proposition.

PROPOSITION 3. *Let $P_\lambda = P_\lambda(\alpha, \beta, \gamma)$ denote the probability that a random filling of the diagram λ with letters from L with independent values and common distribution of type (α, β, γ) produces an L-tableau* (*in the sense of* § 2). *Then $P_\lambda(\alpha, \beta, \gamma) = e_\lambda(\alpha, \beta, \gamma)$, the extended Schur function.*

Proof. For Young diagrams λ, μ with n cells let $K_{\lambda\mu}$ denote the usual Kostka coefficients [9], [11], i.e., the number of Young tableaux (strictly increasing in columns and weakly increasing in rows) of shape λ containing $\mu_1 1$'s, $\mu_2 2$'s, etc. As in the proof of [11, Thm. 6.2] one can show that

$$H_\mu = \sum_{\lambda \vdash n} K_{\lambda\mu} P_\lambda$$

for all $\mu \vdash n$, where $H_\mu = \prod_k P_{(k)}^{\mu_k}$ and (k) denotes the one-row Young diagram of length k. A similar formula holds for the usual symmetric functions and hence for extended symmetric functions, viz.,

$$h_\mu = \sum_\lambda K_{\lambda\mu} e_\lambda.$$

It is easy to show using Proposition 2 that $P_n(\alpha, \beta, \gamma) = h_n(\alpha, \beta, \gamma)$ for $n=1, 2, \cdots$. It follows that $H_\mu = h_\mu$ for $\mu \vdash n$, so $P_\lambda = e_\lambda$ because the matrix $(K_{\lambda\mu})$ is invertible. □

An important fact which connects the theory of symmetric functions and the theory of random tableaux is given by the following theorem.

THEOREM 1. *Let $m_{\alpha,\beta,\gamma}$ be the product-measure with common multiplier d of type (α, β, γ) and consider the Young tableau $t \in T_n$ of shape $\lambda \vdash n$. Then*

$$m_{\alpha,\beta,\gamma}\{w \in L^\infty \colon S_n(w) = t\} = e_\lambda(\alpha, \beta, \gamma),$$

where $e_\lambda(\alpha, \beta, \gamma)$ is the extended Schur function.

Proof. The probability on the left side of the formula depends only on the sequence w_n of the first n letter of w. Because bitabulation $w_n \to (R_n, S_n)$ is a bijection, the sequences w_n with $S_n(w_n) = t$ are in one-to-one correspondence with L-tableaux of shape λ; so the theorem follows from Proposition 3. □

6. Product-measures, central measures and characters of the group $\mathfrak{S}_\infty$. We shall use several definitions and facts from [16]. The Borel measure M on T is called *central* if

$$M\{t\colon t_n = u\} = M\{t\colon t_n = v\}$$

for every two Young tableaux u, v with the same diagram. The central probability measures on T form a simplex, the extremal points of which are indecomposable central measures corresponding to the factor-representations of finite type of the group $\mathfrak{S}_\infty$. This connection and the complete list of indecomposable central measures have been found in [16] with the help of a method based on the ergodic theorem (compare with the method in [15]). E. Thoma had obtained the list of characters of $\mathfrak{S}_\infty$ in 1964 [14] with a quite different method. From Theorem 1 and the results of [16] the main theorem of the paper follows.

THEOREM 2. *The image $M_{\alpha,\beta,\gamma}$ of the product-measure $m_{\alpha,\beta,\gamma}$ under youngization is an indecomposable central measure. Any indecomposable central measure can be obtained in this way. In other words, the transformation of measures induced by youngization is a bijection between the classes of product-measures* (*with respect to the type*) *and indecomposable central measures.*

Let us give some corollaries of this theorem.

7. Generalized Ulam's problem. According to [16, Cor. 2], if we choose a tableau $t \in T$ with respect to the measure $M_{\alpha,\beta,\gamma}$, then the following limits almost always exist:

$$\lim_n \frac{\tilde{r}_k(\Delta_n)}{n} = \sum_{i=1}^k \alpha_i, \qquad \lim_n \frac{\tilde{c}_k(\Delta_n)}{n} = \sum_{j=1}^k \beta_j.$$

Their existence can also be proved with martingale techniques (compare [6]). Together with Proposition 1 this fact leads to a new law of large numbers for sequences of independent random variables.

PROPOSITION 4. *Let d be a measure of type (α, β, γ) on L and $m_{\alpha,\beta,\gamma} = \prod_1^\infty d$ the product-measure on L^∞. For almost all sequences $w \in L^\infty$ with respect to the measure $m_{\alpha,\beta,\gamma}$ the limits*

$$\lim_n \frac{r_k(w_n)}{n} = \sum_{i=1}^k \alpha_i, \qquad \lim_n \frac{c_k(w_n)}{n} = \sum_{i=1}^k \beta_i$$

exist.

If the measure d is continuous $M = M_{0,0,1}$ is the Plancherel measure; a stronger result for this case was obtained in [16], viz., for almost all tableaux $t = (\lambda_1, \lambda_2, \cdots)$ with respect to the measure M the shape of the diagram in a suitable scale tends (as $n \to \infty$) to a curve Ω which was found and described in [16] (see also [8]). Using the shape of this curve the authors [16] (see [8]) have proved that for the length $\tilde{r}_1$ of the first row of the diagram of the tableau $t = (\lambda_1, \lambda_2, \cdots)$ the inequality

$$\lim_n \frac{\tilde{r}_1(\lambda_n)}{\sqrt{n}} \geqq 2$$

holds both a.e. and in the mean with respect to the Plancherel measure. Moreover, it is easy to prove that $M(A_n) \leqq 1/\sqrt{n}$ where $A_n \subset T$ is the set of tableaux with the nth cell in the first row. In [16] we have obtained from this the inverse inequality

$$\lim_n \frac{\tilde{r}_1(\lambda_n)}{n} \leqq 2$$

and hence the complete solution of Ulam's problem as follows.

THEOREM 3. *The maximal length $r_1(w_n)$ of a monotonic subsequence of a sequence w_n of n independent random variables with arbitrary continuous distribution on L grows as $2\sqrt{n}$ with probability one, i.e.,*

$$m_{0,0,1}\left\{w\colon \lim_n \frac{r_1(w_n)}{\sqrt{n}} = 2\right\} = 1.$$

In such a way we can also obtain information about the growth of several rows and columns. Let us give a closely related fact about symmetric functions.

PROPOSITION 5. *For every $\alpha_1, \alpha_2, \cdots; \beta_1, \beta_2, \cdots; \gamma \geqq 0$, $\sum \alpha_i + \sum \beta_i + \gamma = 1$ and $\varepsilon > 0$ we have*

$$\lim_n \sum (\dim \lambda) e_\lambda(\alpha, \beta, \gamma) = 1,$$

where the sum is over the diagrams λ with n cells for which the inequality $|\tilde{r}_1(\lambda)/n - \alpha_1| < \varepsilon$ holds.

These methods allow us to prove a series of new laws of large numbers for nonlinear functionals for random independent variables (as in Ulam's case) and for random Young tableaux, distributed according to one of the central measures; asymptotic formulae for symmetric functions (as in Propositon 5) also follow.

8. Law of large numbers for representations. Let us consider two integer partitions $\kappa\colon k = k_1 + \cdots + k_s$ and $\mu\colon m = m_1 + \cdots + m_t$, $k + m = n$, and the groups $\mathfrak{S}'_\kappa = \mathfrak{S}_{k_1} \times \cdots \times \mathfrak{S}_{k_s}$, $\mathfrak{S}''_\mu = \mathfrak{S}_{m_1} \times \cdots \times \mathfrak{S}_{m_t}$. Let $V_{\kappa,\mu}$ be the representation of $\mathfrak{S}_n$ induced by the linear character id$\times$sign of the subgroup $\mathfrak{S}'_\kappa \times \mathfrak{S}''_\mu \subset \mathfrak{S}_n$. The Littlewood-Richardson rule (see [9, Chap. I.9]) when applied to $V_{\kappa,\mu}$ can be formulated as follows. Let $L = L_e \cup L_0$ be an alphabet with $|L_e| = k$, $|L_0| = m$. The multiplicity of the irreducible representation $\{\lambda\}$ of the group $\mathfrak{S}_n$ with diagram λ in the representation $V_{\kappa,\mu}$ equals the number of L-tableaux with diagram λ in which the multiplicities of the letters from L_e are $k_1, \cdots, k_s$ and those of the letters from L_0 are $m_1, \cdots, m_t$. Let $X = X_{\kappa,\mu} \subset L^n$ be the set of words with the same multiplicities of letters. Using the above rules, we can prove that the dimension of the λ-primary component in $V_{\kappa,\mu}$ is equal to $|\{x \in X_{\kappa,\mu}\colon \Delta(x) = \lambda\}|$ and the character $\psi = \psi_{\kappa,\mu}$ of the representation $V = V_{\kappa,\mu}$ is the following:

$$\psi(\sigma) = \frac{\dim V}{|X|} \sum_{w \in X} \chi_{\Delta(w)}(\sigma)/\dim \Delta(w)$$

where $\sigma \in \mathfrak{S}_n$ and χ_λ is the character of the representation $\{\lambda\}$.

Applying these facts to our case, together with Theorems 1 and 2, we get:

PROPOSITION 6. *Let $\psi^{(n)}$ be the character of the induced representation $V^{(n)}$ with the parameters $\kappa(n)$, $\mu(n)$ (see above) and*

$$\lim_n \frac{k_i(n)}{k(n) + m(n)} = \alpha_i, \qquad \lim_n \frac{m_i(n)}{k(n) + m(n)} = \beta_i.$$

Then the limit

$$\lim_n \frac{\psi^{(n)}(\sigma)}{\dim V^{(n)}} = \chi_{\alpha,\beta,\gamma}(\sigma), \qquad \sigma \in \mathfrak{S}_\infty$$

gives the primary character of the group $\mathfrak{S}_\infty$ corresponding to the central measure $M_{\alpha,\beta,\gamma}$. Thus, the limit of the induced characters is the same as the limit of primary (irreducible) characters with the same parameters.

Denote by $\mathscr{D}_{\kappa,\mu}(\lambda)$ the relative dimension of the primary component, corresponding to the diagram $\lambda \vdash n$ in the expansion of $V_{\kappa,\mu}$.

PROPOSITION 7. *Under the conditions of Proposition* 6, *for any* $\varepsilon > 0$ *and* $k = 1, 2, \cdots$, *we have*

$$\lim \mathscr{D}_{\kappa,\mu} \left\{ \lambda : \left| \frac{\tilde{r}_k(\lambda)}{n} - \sum_{i=1}^{i} \alpha_i \right| \geqq \varepsilon, \left| \frac{c_k(\lambda)}{n} - \sum_{i=1}^{k} \beta_i \right| \geqq \varepsilon \right\} = 0.$$

In other words, for large n and for a typical diagram of an irreducible representation in the expansion of $V^{(n)}$, the relative length of the ith row (column) is close to $k_i/n(m_i/n)$.

These properties can be called the *laws of large numbers for characters and representations* because they express how (in a statistical sense) the irreducible components of the induced representation concentrate (for large n) near one of them. Probably this property takes place for a wider class of locally finite groups. We can formulate it as a thesis: The list of the limits of the primary characters coincides with that of the limits of characters induced from a suitable class of subgroups (in the $\mathfrak{S}_\infty$-case, Young subgroups).

9. Analogues of the main theorem for other graphs. Let $\mathscr{D}$ be a branching scheme (Bratteli diagram), i.e., a graded graph with the set $\mathscr{D}^v = \bigcup_{n=0}^{\infty} \mathscr{D}_n$ of vertices; the edges (possibly multiple) join the vertices of neighbouring levels only. A *path* in this graph $\mathscr{D}$ is a sequence $t = (e_1, e_2, \cdots)$ of edges in which e_{n-1} and e_n have a common vertex in $\mathscr{D}_n$, $n = 1, 2, \cdots$. The space of paths $\operatorname{Ch} \mathscr{D} = \varprojlim \operatorname{Ch} \mathscr{D}_n$ is compact. A Borel measure on $\operatorname{Ch} \mathscr{D}$ is by definition a *central measure* if the condition of § 6 holds with u and v being finite paths (elements of $\operatorname{Ch}_n \mathscr{D}$) with a common end; t_n is the initial segment of length n of the path $t \in \operatorname{Ch} \mathscr{D}$.

The problem of describing the indecomposable central measures for branching schemes is the most interesting problem in the theory of the representations of AF-algebras and locally finite groups (for factor-representations and K-functors see [5]).

For some graded graphs $\mathscr{D}$ it is possible to find the space $\mathscr{F}$ and a cylindrical transformation ϕ of the infinite product $\mathscr{F}^\infty$ onto $\operatorname{Ch} \mathscr{D}$, $\phi : \mathscr{F}^\infty \to \operatorname{Ch} \mathscr{D}$, such that the corresponding transformation of measures gives a surjection of the space of product-measures onto that of indecomposable central measures. We shall call this kind of graph $\mathscr{D}$ *projective* and the above transformation ϕ its *projectivization.* The basic example is the Young graph.

Example 1. The vertices of the graph $\mathscr{D} = Y$ are Young diagrams, the grading is defined by the number of cells, and the edges are as usual (i.e., two diagrams λ and μ are adjacent if λ can be obtained from μ by adjoining a single cell). Then $\operatorname{Ch} \mathscr{D} = T$ is the space of infinite Young tableaux. We can take for $\mathscr{F}$ the alphabet L from §§ 2–6. Theorem 2 asserts that Y is projective and the youngization is its projectivization.

A completely different example is contained in a nonevident way in [6].

Example 2 (Kingman's graph). The set of vertices is as in the first example, but we use another language. The nth level in the graph K is the set of all (nonordered) partitions of the integer n, $n = n_1 + \cdots + n_k$. The edges are defined as follows. Consider two partitions $\lambda \in K_n$ and $\Lambda \in K_{n+1}$. They are joined in K iff Λ has the same parts as λ, excluding only one part n_i. In this case there is a one-to-one correspondence between (a) the set of edges joining λ and Λ, and (b) parts of Λ equal to n_i. Thus we can identify every edge entering Λ with a part of Λ. Let $\lambda_n(w)$ be the partition of n into the multiplicities of the letters which one meets in the initial piece w_n of the sequence $w \in \mathscr{F}^\infty$ ($\mathscr{F}$ is an alphabet). The last letter x_n in w_n fixes some part in the partition

$\lambda_n(w)$. For every $w \in \mathscr{F}^\infty$ we get the sequence of edges which is the path $\phi(w) \in \mathrm{Ch}\, K$. If we put $\mathscr{F} = [0, 1]$, the map $\phi: \mathscr{F}^\infty \to \mathrm{Ch}\, K$ gives the projectivization of K. This fact can be extracted from [6], where the list of central measures (called there "partition structures") was found. It is interesting that the central measures in this case have a connection with the asymptotic theory of Haar measure on $\mathfrak{S}_n$ (see [18]). Here is an explanation of this fact: the graph K is that of conjugacy classes of the groups $\mathfrak{S}_n$, with edges which are defined naturally by the imbedding $\mathfrak{S}_n \subset \mathfrak{S}_{n+1}$.

Example 3 (the order ideals of the universal binary tree). This example arises in connection with [12]. Let $\mathscr{D}$ be the Hasse diagram of the lattice of finite order ideals of the universal binary rooted tree T_2. Let $\mathscr{F}$ be the space of infinite chains in T_2; $\mathscr{F} \cong \prod_1^\infty \mathbb{Z}_2$. For $w = (w_1, w_2, \cdots) \in \mathscr{F}^\infty$ we put $\phi(w) = (d_0, d_1, \cdots) \in \mathrm{Ch}\, \mathscr{D}$, where $d_0 = \varnothing$, $d_n = d_{n-1} \cup a$, $a \in T_2$ is the vertex of T_2 in which the chain w_n leaves the ideal d_{n-1}. It can be shown that ϕ is the projectivization of the graph $\mathscr{D}$.

In conclusion we note that the problem of describing the central measures of a graded graph can be formulated as the problem of describing the set of Markovian measures on the compact space of paths of this graph with given cotransition probabilities (which are common for all central measures). So this problem is included in the cycle of questions which are typical of modern statistical physics.

Comment. This paper was offered to the journal at the end of 1981. For reasons not depending on the authors, its publication was postponed. During this time we were informed about the articles [1], [10], [13], [19], where the generalized Schur functions are introduced in connection with a quite different approach (Lie superalgebras). However these functions were considered for the first time in our article [20] as the characters of factor-representations of the group $\mathfrak{S}_n$.

In the recent paper [21] following the paper [17], we have obtained an exact asymptotical estimation of the maximal degree of irreducible representation of $\mathfrak{S}_n$:

$$0 < C_0 \leqq -\frac{1}{\sqrt{n}} \ln \max \frac{\dim \Lambda}{\sqrt{n!}} \leqq C_1.$$

Acknowledgment. The authors thank professor R. P. Stanley for his attention to the paper, for linguistic help and for the references [1], [10], [13].

REFERENCES

[1] A. Berele and A. Regev, *Hook Young diagrams with applications to combinatorics and to representations of Lie superalgebras*, preprint.

[2] D. Foata, ed., *Combinatoire et représentation du groupe symétrique*, Lecture Notes in Mathematics. 579, Springer-Verlag, Berlin/Heidelberg/New York, 1977.

[3] S. V. Fomin, *Finite partially ordered, sets and Young diagrams*, Dokl. Akad. Nauk USSR, 243 (1978), pp. 1144-1147, Soviet Math. Dokl., 19 (1978), pp. 1510-1514.

[4] C. Greene and D. J. Kleitman, *On the structure of Sperner k-families*, J. Combin. Theory (A), 20 (1976), pp. 41-68.

[5] S. V. Kerov and A. M. Vershik, *Characters, factor-representations and K-functor of the infinite symmetric group*, Operator Algebras and Group Representatives, Vol. II, Monographs and Studies in Mathematics, Pitman, London, 1984, pp. 23-32.

[6] J. F. C. Kingman, *The representation of partition structures*, J. London Math. Soc., (2), 18 (1978), pp. 374-380.

[7] D. E. Knuth, *Permutations, matrices, and generalized Young tableaux*, Pacific J. Math., 34 (1970), pp. 709-727.

[8] B. F. Logan and L. A. Shepp, *A variational problem for random Young tableaux*, Advances in Math., 26 (1977), pp. 206-222.

[9] I. G. Macdonald, *Symmetric Functions and Hall Polynomials*, Oxford Univ. Press, Oxford, 1979.

[10] J. B. REMMEL, *The combinatorics of (k, l)-hook Schur functions*, Contemporary Math., 34 (1984), to appear.

[11] R. P. STANLEY, *Theory and application of plane partitions, Part* I, Studies in Applied Math., 50 (1971), 167-188.

[12] ———, *The Fibonacci lattice*, Fibonacci Quart., 15 (1975), pp. 215-232.

[13] ———, *Unimodality and Lie superalgebras*, Studies in Applied Math., to appear.

[14] E. THOMA, *Die unzerlegbaren, positiv—definiten Klassenfunktionen der abzählbar unendlichen, symmetrischen Gruppe*, Math. Z., 85 (1964), pp. 40-61.

[15] A. M. VERSHIK, *Description of invariant measures for the actions of some infinite-dimensional groups*, Dokl. Akad. Nauk. USSR, 218 (1974), pp. 749-752; Soviet Math. Dokl., 15 (1974), pp. 1396-1400.

[16] A. M. VERSHIK AND S. V. KEROV, *Asymptotic theory of characters of the symmetric group*, Funct. Analis i Pril., 15 (1981), pp. 17-27; Funct. Anal. Appl., 15 (1981), pp. 246-255.

[17] ———, *Asymptotics of the Plancherel measure of the symmetric group and the limiting form of Young tableaux*, Dokl. Akad. Nauk. USSR, 233 (1977), 1024-1027; Soviet Math. Dokl., 18 (1977), pp. 527-531.

[18] A. M. VERSHIK AND A. A. SCHMIDT, *Limiting measures arising in the asymptotic theory of symmetric groups*, Teor. Ver. i Pril., 22 (1977), pp. 72-88; 23 (1978), pp. 42-54; Theory Prob. Appl., 22 (1977), pp. 70-85; 23 (1978), pp. 36-49.

[19] A. N. SERGEEV, *Tensor algebra of the identity representation as a module over the Lie superalgebra $GL(n, m)$, $Q(h)$*, Mat. Sb., 123, 3 (1984), pp. 422-430. (In Russian.)

[20] A. M. VERSHIK AND S. V. KEROV, *Characters and factor representations of the infinite symmetric group*, Dokl. Akad. Nauk. SSSR, 275, 5 (1981); Soviet Math. Dokl., 23, 2 (1981), pp. 389-397.

[21] ———, *Asymptotics of maximal and typical dimensions of the irreducible representations of the symmetric group*, Funct. Anal., 19, 1 (1985), pp. 25-36. (In Russian.)

[22] ———, *K-functor (Grothendic group) of the infinite symmetric group*, Zap. Nauk. Sem. LOMI (Leningrad), 123 (1983), pp. 126-151 (in Russian); *The Representations of Infinite Dimensional Lie Algebras*, Gordon and Breach, New York, 1985.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

THE NUMBER OF MAXIMAL INDEPENDENT SETS IN A TREE*

HERBERT S. WILF†

Abstract. We find the largest number of maximal independent sets of vertices that any tree of n vertices can have.

AMS(MOS) subject classifications. 05C05, 05C15, 05C35

Key words. maximal independent sets, cliques, tree

1. Introduction. We determine, in § 2 below, the largest number of maximal independent sets of vertices that any tree of n vertices can have. In § 3 there is a linear time algorithm for the computation of the number of maximal independent sets of any given tree. The application that suggested these questions to us was the analysis of the complexity of an algorithm for computing the chromatic number of a graph. That application will be discussed in § 4.

2. The main theorem. Let T be a tree, let $V(T)$ be its vertex set and let $n=|V(T)|$ be its number of vertices. A set $S\subseteq V(T)$ is an *independent set* if no two vertices of S are joined by an edge of T. S is a *maximal independent set* (*m.i.s.*) if S is independent and every vertex of $V(T)-S$ is joined by an edge to at least one vertex of S. We write $\mu(T)$ for the number of m.i.s. of vertices of T ($\mu(\varnothing)=1$).

THEOREM 1. *If we define*

$$f(n)=\begin{cases}2^{n/2-1}+1 & \text{if } n\geqq 2 \text{ is even,}\\ 2^{(n-1)/2} & \text{if } n \text{ is odd,}\\ 1 & \text{if } n=0,\end{cases}\tag{1}$$

then $f(n)$ is the largest number of maximal independent sets of vertices that any tree of n vertices can have.

Figure 1 shows that there are trees of n vertices that have $f(n)$ maximal independent sets (the reader may enjoy checking these counts since they are not quite trivial!). Hence it suffices to prove that no n-tree can have more than $f(n)$ such sets.

Let T be a tree of $n\geqq 3$ vertices, and let x be an endpoint of T. We *root* T at x and direct the edges of T away from x.

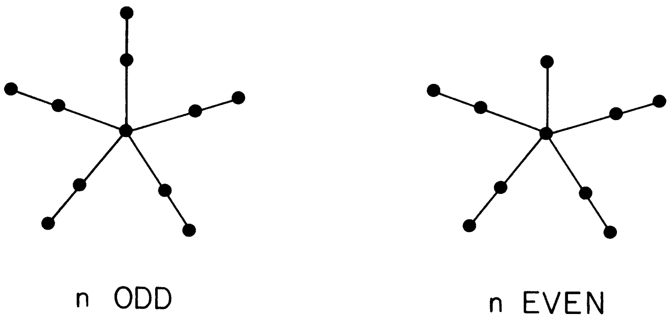


FIG. 1

* Received by the editors February 10, 1984, and in revised form October 15, 1984.

† Department of Mathematics, University of Pennsylvania, Philadelphia, Pennsylvania 19104.

Let $\gamma=\gamma(x)$ be the child of x and let $\lambda_1, \cdots, \lambda_r$ be the children of γ. Let U_i be the subtree of T that is rooted at λ_i $(i=1, r)$.

We continue one layer further into T: in U_i, let $W_{i,j}$ $(j=1, s_i)$ be the subtrees that are rooted at the s_i children of λ_i, except that if λ_i is childless then we take $s_i=1$, and $W_{i,1}$ is then the empty tree $(i=1, r)$. The picture is now as shown in Fig. 2.

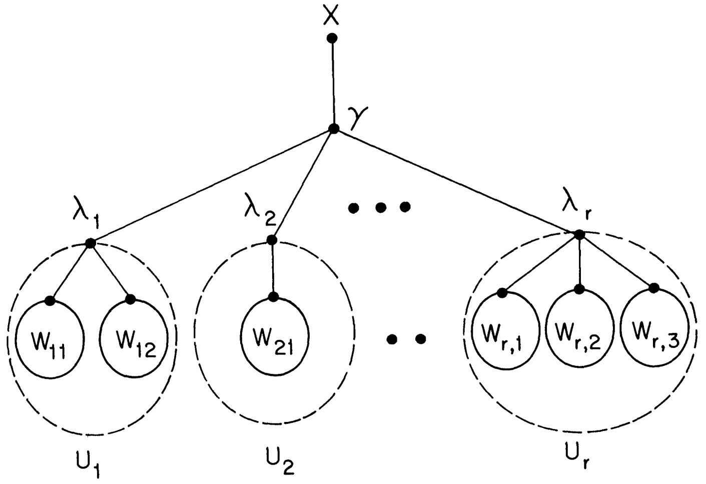


FIG. 2

LEMMA 1. *If T is a tree of $n \geqq 3$ vertices then*

$$\mu(T)=\prod_{i=1}^{r} \mu(U_i)+\prod_{i=1}^{r} \prod_{j=1}^{s_i} \mu(W_{i,j}). \tag{2}$$

Proof. Let $S \subseteq V(T)$ be a m.i.s. that contains x. Then $\gamma \notin S$. Let $S_i=S \cap V(U_i)$ $(i=1, r)$. Then S_i is maximal in U_i $(i=1, r)$, for if not then S can be augmented in T. Conversely, if $\forall i=1, r$: S_i is maximal in U_i, then $S=\{x\} \cup S_1 \cup \cdots \cup S_r$ is maximal in T.

Next consider a m.i.s. $S \subseteq V(T)$ such that $x \notin S$, and therefore $\gamma \in S$. Hence $\forall i=1, r$: $\{\lambda_i \notin S$, and $\forall j=1, s_i$: {if $S_{i,j}=S \cap V(W_{i,j})$ then $S_{i,j}$ is maximal in $W_{i,j}$ and conversely}}. □

Let

$$h(n)=\max_{|V(T)|=n} \mu(T). \tag{3}$$

We will now prove that $\forall n \geqq 0$: $h(n)=f(n)$. Clearly $h(n)=f(n)$ if $n \leqq 2$. Suppose that $n \geqq 3$, and that $\forall j=0, n-1$: $h(j)=f(j)$. Let T be a tree of n vertices, and let x, $U_i(T)$, $W_{i,j}(T)$ be as in Fig. 2. Write $u_i=|V(U_i(T))|$ $(i=1, r)$ and $w_{i,j}=|V(W_{i,j}(T))|$ $(j=1, s_i; i=1, r)$. Then by (2) and the induction hypothesis,

$$\mu(T) \leqq \prod_{i=1}^{r} f(u_i)+\prod_{i=1}^{r} \prod_{j=1}^{s_i} f(w_{i,j}). \tag{4}$$

We will carry out a maximization of (4) over all n-trees T in two stages, as follows. As the problem is presented in (4) we are to maximize the right-hand side over all partitions $\mathbf{u}$ of $n-2$ and all partitions $\mathbf{w}$ of the parts of $\mathbf{u}$ (each reduced by 1). In Stage 1 below we will identify, for given $\mathbf{u}$, the maximizing partition $\mathbf{w}$, and we will be left with maximization over just the partitions $\mathbf{u}$.

In Stage 2 we will show that the maximum depends only on two integers, the number r of parts of $\mathbf{u}$, and the number e of *even* parts of $\mathbf{u}$, but not otherwise on $\mathbf{u}$. We will then carry out the maximization over the admissible integer pairs (r, e), with the end result that the maximum of the right side of (4) will have been shown to be $f(n)$, as defined in (1).

Stage 1 (*in which the trees and the $w_{i,j}$'s are eliminated*).

Fix integers $m, r \geqq 1$, let $\Gamma(r, m)$ denote the set of all r-tuples of positive integers whose sum is m, and write $\Gamma(1, 0) = \{0\}$. If we take the maximum of (4) over all n-trees T, we get

$$(5) \qquad h(n) \leqq \max_{r \geqq 1} \max_{\mathbf{u} \in \Gamma(r, n-2)} \left\{ \prod_{i=1}^{r} f(u_i) + \max \prod_{i=1}^{r} \prod_{j=1}^{s_i} f(w_{i,j}) \right\}$$

in which the innermost "max" is over the set of $\mathbf{w}$'s such that for $i = 1, r$:

$$(w_{i,1}, \cdots, w_{i,s_i}) \in \Gamma(s_i, u_i - 1).$$

Consider an integer r and partitions $\mathbf{u}$, $\mathbf{w}$ that occur on the right side of (5) and in which one or more of the $w_{i,j} \geqq 3$. We claim that this set of integers can be ignored when seeking the maximum in (5).

Indeed, replace $w_{1,1}$ by $\lfloor w_{1,1}/2 \rfloor$ 2's, plus, possibly a 1, leaving all other w's, u's and r untouched. Then the double product on the right will contain a factor

$$f(2)^{\lfloor w_{1,1}/2 \rfloor} = 2^{\lfloor w_{1,1}/2 \rfloor}$$

instead of the factor $f(w_{1,1})$. But from (1), $f(k) \leqq 2^{\lfloor k/2 \rfloor}$ for all $k \geqq 3$. Hence the right side of (5) cannot decrease by such a replacement.

Therefore, for fixed r and $\mathbf{u} \in \Gamma(r, n-2)$ we need consider only partitions of each $u_i - 1$ into 0's, 1's and 2's, say α 2's, β 1's and γ 0's $(\alpha \leqq (u_i - 1)/2)$. However, such a partition of $u_i - 1$ contributes a factor of 2^α to the innermost product in (5), and this is maximal when $\alpha = \lfloor (u_i - 1)/2 \rfloor$. Hence for r fixed and $\mathbf{u} \in \Gamma(r, n-2)$, the double product in (5) cannot exceed

$$\prod_{i=1}^{r} 2^{\lfloor (u_i - 1)/2 \rfloor} = 2^{(n-2-r-e)/2}$$

where $e = e(\mathbf{u})$ is the number of even numbers among $u_1, \cdots, u_r$.

Stage 2 (*in which the u_i's are eliminated*).

As a result of stage 1 we have found that

$$(6) \qquad h(n) \leqq \max_{r \geqq 1} \max_{\mathbf{u} \in \Gamma(r, n-2)} \left\{ \prod_{i=1}^{r} f(u_i) + 2^{(n-2-r-e(\mathbf{u}))/2} \right\}.$$

Thus we have now a maximization problem over integer partitions, instead of over trees.

Fix three integers r, e, t such that $r \geqq 1$, $0 \leqq e \leqq r$ and $t \geqq e$. Consider the subclass $J(r, e, t) \subseteq \Gamma(r, n-2)$ of those partitions of $n-2$ into r positive parts, exactly e of which are even, and in which the sum of the even parts is $2t$. More precisely, then, $J(r, e, t)$ is the class of all partitions of the form

$$\begin{aligned} &\text{(a)} \quad n-2 = 2l_1 + 2l_2 + \cdots + 2l_e + (2l_{e+1} + 1) + \cdots + (2l_r + 1), \\ (7) \quad &\text{(b)} \quad l_i \geqq 1 \quad (i = 1, e), \qquad l_i \geqq 0 \quad (i = e+1, r), \\ &\text{(c)} \quad l_i + \cdots + l_e = t. \end{aligned}$$

Among the partitions $\mathbf{u} \in J(r, e, t)$ the second term in the brace in (6) is constant, so we consider

$$\begin{aligned}\max_{J(r,e,t)} \prod_{i=1}^{r} f(u_i) &= \max_{J(r,e,t)} \prod_{i=1}^{e} f(2l_i) \prod_{i=e+1}^{r} f(2l_i+1)\\ &= \max_{J(r,e,t)} \prod_{i=1}^{e} (2^{l_i-1}+1) \prod_{i=e+1}^{r} 2^{l_i}\\ &= 2^{(n-2-r-e)/2} \max_{J(r,e,t)} \prod_{i=1}^{e} (1+2^{-(l_i-1)})\end{aligned}\tag{8}$$

where (7a) was used in the last equality.

The following result will be useful in the sequel.

LEMMA 2. *Fix $g, z \geqq 0$. Let*

$$G(g, z) = \max \prod_{i=1}^{g} (1+2^{-m_i})\tag{9}$$

where the max is over all g-tuples of nonnegative integers $m_1, \cdots, m_g$ whose sum is z. Then

$$G(g, z) = 2^{g-1}(1+2^{-z})\tag{10}$$

and the maximum occurs when exactly one $m_i = z$ and all $m_j = 0$ $(j \neq i)$. □

It is now convenient to split up the maximization of (8) over $J(r, e, t)$ into two cases, first where $e = r$, so all parts of (7)(a) are even, and second where $e < r$, so odd parts then also occur in (7)(a).

Case I. $e = r$. In this case (7a) shows that n is even and

$$\sum_{i=1}^{r} l_i = n/2-1 = t$$

so we are in the class $J(e, e, n/2-1)$. If we use Lemma 2 with $g = e$, $z = n/2-1-e$, the maximum on the right side of (8) becomes just $2^{n/2-2}+2^{e-1}$. In this case, then, (6) takes the form

$$h(n) \leqq \max_{e} \{2^{n/2-2}+2^{e-1}+2^{n/2-1-e}\}\tag{11}$$

where the maximum extends over $1 \leqq e \leqq n/2-1$. It is clear, from (11), that the maximum occurs at either endpoint $e = 1$ or $e = n/2-1$, of the interval. The maximum value is $1+2^{n/2-1} = f(n)$, as required.

Case II. $e < r$. Here we find, from Lemma 2 with $g = e$, $z = t-e$, that the maximum on the right side of (8) is

$$\max_{J(r,e,t)} \{2^{(n-2-r-e)/2}\{2^{e-1}+2^{2e-t-1}\}\}.\tag{12}$$

The maximum over t occurs when t is as small as possible, viz. $t = e$ (see (7b, c)) and the maximum is $2^{(n-2-r+e)/2}$. Thus (6) now becomes

$$h(n) \leqq \max_{(r,e)} \{2^{(n-2-r+e)/2}+2^{(n-2-r-e)/2}\}.\tag{13}$$

In (13) the "max" is taken over the set of (r, e) for which

$$(14)\quad \begin{array}{lll} \text{(a)} & 1 \leqq r \leqq n-2 & \text{(from (7a))} \\ \text{(b)} & 0 \leqq e < r & \text{(in Case II)} \\ \text{(c)} & e+r \leqq n-2 & \text{(from (7a, b))} \\ \text{(d)} & e+r \equiv n (\operatorname{mod} 2) & \text{(from 7a)} \end{array}$$

Suppose n is odd. We claim that the "max" in (13) occurs at $r=1$, $e=0$. Indeed, if it occurs at (r, e) then surely $r=e+1$ or $r=e+2$, else we could reduce r by 2 to increase the maximum without violating any of the constraints (14). Hence $r=e+1$, by (14d), and (13) reads as

$$h(n) \leqq 2^{n/2-3/2} \max_e \{1+2^{-e}\} = 2^{(n-1)/2} = f(n)$$

so in Case II, n odd, we have established that $h(n) \leqq f(n)$.

Finally, in Case II, suppose n is even, and further suppose that the maximum in (13) occurs at (r, e). Again we must have $r=e+1$ or $r=e+2$, else we could reduce r by 2. Now $r=e+1$ is ruled out by (14d), so $r=e+2$. Therefore (13) reduces to

$$(15)\qquad h(n) \leqq 2^{n/2-2} \max_e \{1+2^{-e}\}$$

and the max occurs at $e=0$, the value being $2^{n/2-1} < f(n)$, completing the proof of Theorem 1. □

3. A linear time algorithm. In this section we give another algorithm for computing $\mu(T)$. It will easily be seen to operate in linear time.

Let the edges of T be oriented away from the root r, let x be some vertex, and let $\mathscr{C}(x)$, $\mathscr{G}(x)$ be the sets of children of x and of grandchildren of x, respectively. Let μ_x be the number of m.i.s. in the subtree rooted at x, and let ν_x be the number of those m.i.s. that do not contain x. Then it is easy to see that

$$(16)\qquad \nu_x = \prod_{y \in \mathscr{C}(x)} \mu_y - \prod_{y \in \mathscr{C}(x)} \nu_y, \qquad \mu_x = \nu_x + \prod_{z \in \mathscr{G}(x)} \mu_z .$$

These formulas permit the computation of the pairs (μ_x, ν_x) at each vertex of T, in descending order of distance from r. One would begin by introducing a new fictitious "child" of each leaf, and placing $(1, 0)$ at each such new vertex as well as at each leaf. The remaining vertices could then be done, in descending order, from (16). Therefore the number of maximal independent sets of vertices in a tree can be computed in linear time.

4. Remarks. In [1] E. Lawler discusses an algorithm for determining the chromatic number of a graph, and shows that its run time, in the worst case, is $O(mn(1+\sqrt[3]{3})^n)$ for graphs of m edges and n vertices.

The appearance of $\sqrt[3]{3}$ derives from a theorem of Moon and Moser [2] to the effect that a graph of n vertices cannot have more than $3^{n/3}$ maximal independent sets (they proved a sharper bound, but this one suffices for our present purpose). However, the extremal graphs of Moon and Moser are disconnected. They are essentially disjoint unions of triangles.

An improvement of Lawler's run time estimate might therefore result if we could solve the following problem:

What is the largest number of maximal independent sets that can occur in a connected graph of n vertices?

The present paper resulted from consideration of the above question. J. Griggs, C. Grinstead and D. Guichard (p.c.) have shown that if $c(n)$ denotes the answer to this question then $\lim c(n)^{1/n} = 3^{1/3}$. *Note added in proof.* They and, independently, Z. Füredi, have now answered the above question.

REFERENCES

[1] E. LAWLER, *A note on the complexity of the chromatic number problem*, Inform. Proc. Lett., 5 (1976), pp. 66–67.

[2] J. MOON AND L. MOSER, *On cliques in graphs*, Israel J. Math., 3 (1965), pp. 23–28.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

016

EFFICIENT VERTEX- AND EDGE-COLORING OF OUTERPLANAR GRAPHS*

ANDRZEJ PROSKUROWSKI† AND MACIEJ M. SYSŁO‡

Abstract. The problems of finding values of the chromatic number and the chromatic index of a graph are NP-hard even for some restricted classes of graphs. Every outerplanar graph has an associated tree structure which facilitates algorithmic treatment. Using that structure, we give an efficient algorithm to color the vertices of an outerplanar graph with the minimum number of colors. We also establish algorithmically the value of the chromatic index of an outerplanar graph. Our algorithms are based on systematic coloring of elements (vertices and edges, respectively) of adjacent faces.

AMS(MOS) subject classifications. 05C15, 05C05, 68Q25

1. Introduction. The *chromatic number*, $\chi(G)$, of a graph G is the minimum number of colors needed for the vertices of G so that no two adjacent vertices are assigned the same color. Correspondingly, the *chromatic index*, $\chi'(G)$, is defined as the minimum number of colors needed to color the edges of G so that no two adjacent edges are assigned the same color. An assignment of at most k colors to the vertices (edges) of a graph G is called a *k-vertex-* (*k-edge-*) *coloring* of G. The problem of determining the chromatic number (index) of a graph G is NP-complete, even when G is restricted to be planar (respectively 3-valent, see Garey and Johnson [6], Holyer [8]). However, Gabow and Kariv [5] have designed an efficient edge-coloring algorithm for bipartite graphs, Mitchell and Hedetniemi [9] have a linear algorithm for edge-coloring trees and unicyclic graphs, and recently, Widgerson [14] has presented an efficient approximation algorithm for vertex-coloring general graphs. Arjomandi [2] and Terada and Nishizeki [13] present approximate algorithms for edge-coloring general graphs. Applying a method that follows the recursive construction of series-parallel graphs, we can easily color vertices of a series-parallel graph using the minimum number of colors in time proportional to the size of the graph. (Compare with, for instance, Takamizawa et al. [12] who do not treat explicitly the problems of vertex- or edge-coloring of graphs, however.)

Here, we present efficient algorithms vertex- and edge-coloring graphs of the subclass of series-parallel graphs, known as *outerplanar graphs.* A planar graph G is outerplanar if and only if there exists a plane embedding of G in which all vertices lie on the exterior (unbounded) face. Such an embedding is referred to as an *outerplane graph.* For every 2-connected outerplane graph G there is a unique *associated tree* $T(G)$. This tree has internal nodes corresponding to the interior (bounded) faces of G, and external nodes (leaves) corresponding to the exterior face, one leaf for each edge of G on the exterior face. This associated tree corresponds to the—possibly cyclic—weak dual graph of a general plane graph. To avoid confusion, we will talk about *nodes* of $T(G)$ and *vertices* of G. The edges of $T(G)$ correspond uniquely to edges of G in such a way that there is an edge between nodes of $T(G)$ if and only if the two corresponding faces of G share an edge. We consider $T(G)$ to be a plane

* Received by the editors July 11, 1983, and in revised form October 18, 1984.

† College of Business Administration, Creighton University, Omaha, Nebraska 68178; on leave from Department of Computer and Information Science, University of Oregon, Eugene, Oregon 97403. The work of this author was partially supported by the National Science Foundation under grant NSF-INT-83 18441.

‡ Department of Computer Science, Washington State University, Pullman, Washington 991645; on leave from Institute of Computer Science, University of Wrocław, Wrocław, Poland.

tree, in which the neighborhood of each node is ordered (see Proskurowski and Sysło [11]). A choice of a node of $T(G)$ as its *root* induces a natural *father-son* relation between adjacent nodes, and also a left-to-right ordering of *brother* nodes (sons of a common father). A corresponding structure for a separable outerplane graph G is the associated forest $F(G)$. In that case, the connected components of $F(G)$ correspond to blocks (2-connected components) of G. The tree-like structure of these blocks allow an easy "color-exchange" algorithm transforming partial solutions for blocks of G into a global solution, an optimal vertex- or edge-coloring of G. Thus, without loss of generality, we restrict our discussion of coloring vertices and edges to 2-connected outerplane graphs. We may assume that a linear time algorithm has been used to extract 2-connected components from a general outerplane graph. See Fig. 1 for an example of an outerplane graph, its associated tree and a rooting.

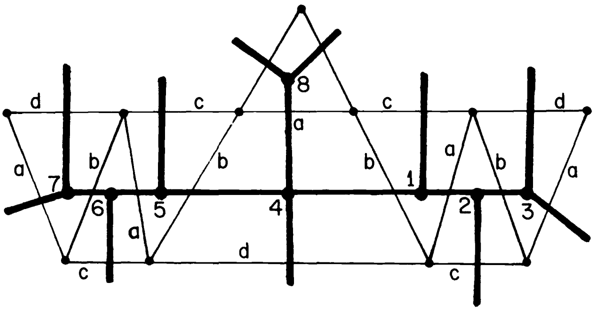


FIG. 1. *An outerplane graph, its associated tree (rooted at node* 1) *with a depth-first traversal order, and a partial edge-coloring following that order.*

In the remainder of this paper we follow the standard texts of Fiorini and Wilson [4], Garey and Johnson [6], and Harary [7] as references for edge-coloring, complexity analysis, and general graph theory, respectively.

2. Vertex-coloring. The Four Color Theorem (Appel and Haken [1]) ensures that the chromatic number of any outerplanar graph (as a planar graph) is at most 4. The fact that the chromatic number of an outerplanar graph is at most 3 is implied by the following observation. Every outerplanar graph has a vertex of degree 2. Every subgraph of an outerplanar graph is outerplanar. Hence, we can apply the Szekeres–Wilf bound on the chromatic number $\chi(G) \leqq 1 + \max \delta(G')$, where maximum is over all subgraphs G' of G, and $\delta(G')$ is the minimum vertex degree of G'.

THEOREM 1. *The chromatic number of an outerplanar graph is at most* 3.

Although the same result can be obtained using the techniques of Takamizawa et al. [12], we enclose our new algorithm for completeness of the presentation.

Our method for producing an optimal vertex-coloring of a 2-connected, outerplane graph G makes use of a traversal of the associated tree $T(G)$, rooted at an arbitrary node. We assume that the traversal is *monotonic*, that is, no node other than the root is visited before its father.

Visiting a node C of $T(G)$ we color the vertices of the corresponding face of G with two or three colors, depending on its length. If C is not the root, two of its adjacent vertices are already colored. These colors are subsequently used to color the cycle C. It is clear that an outerplanar graph containing an odd-length face is not bipartite. Our algorithm will produce a 3-coloring of such a graph. If all faces of G have an even length then a 2-coloring is produced.

This algorithm for coloring the vertices of an outerplane graph G takes time linear in the total size of all faces of G, which, in turn, is proportional to the number of vertices in G.

3. Breadth-first edge-coloring algorithms. The chromatic index of a graph G is bounded by the maximum degree, $\Delta(G)$, of a vertex of G. Vizing's theorem (see, for instance, Fiorini and Wilson [4]) states that $\Delta(G) \leqq \chi'(G) \leqq 1 + \Delta(G)$. Fiorini [3] proves that for an outerplanar graph G, $\chi'(G) = \Delta(G)$ unless G is an odd cycle. However, his proof is an existential one and does not provide a method for finding an optimal edge-coloring. Another proof of the above equality given in Fiorini and Wilson [4] contains a flaw. Even corrected, their proof yields an edge-coloring algorithm of higher-than-linear complexity. Terada and Nishizeki [13] also prove this result. We present an algorithm for optimally edge-coloring an outerplanar graph, which may be consided as yet another, constructive proof of the above equality. The algorithm takes explicit advantage of the structure of a 2-connected outerplane graph by assigning edge colors as it traverses the associated tree of the graph. Our algorithm has time complexity linear with the size of the input, while the exact algorithm of Nishizeki et al. [10] for edge-coloring series-parallel graphs is less efficient, as is the approximate method of Arjomandi [2]. (They have complexity of $O(mn)$ and $O(\min(mn, n\Delta + m\sqrt{n}\log n))$, respectively, where m and n are orders of the graph's edge and vertex sets.)

The arbitrary monotonic traversal of the arbitrarily rooted associated tree $T(G)$, used above in an optimal vertex-coloring of an outerplanar graph G fails in an attempt to edge-color G. The free choice of coloring edges along a cycle, when restricted by an algorithmic method may lead to an eventual coloring conflict. See Fig. 1, where a monotonic traversal is used. The edges of the triangular face corresponding to node 8 cannot be colored without the use of a fifth color.

We will give a traversal method of a carefully rooted associated tree $T(G)$, and a judicial coloring of the corresponding cycles of G that produces an optimal edge-coloring. We define a *breadth-first traversal* of the internal nodes of a rooted plane tree T as the traversal of the nodes of T in left-to-right order in levels defined by the distance from the root. Figure 2 indicates the order of node visits in the breadth-first traversal of a rooted plane tree. The process of visiting a node E during the traversal of $T(G)$ corresponds to edge-coloring the corresponding face E of G. Although at most one edge e of E has a color already assigned (during a visit of the node's father, C), the color assignment to the two edges of E adjacent to e is restricted by other colored edges adjacent to e. This restriction may prevent an optimal coloring if, in the

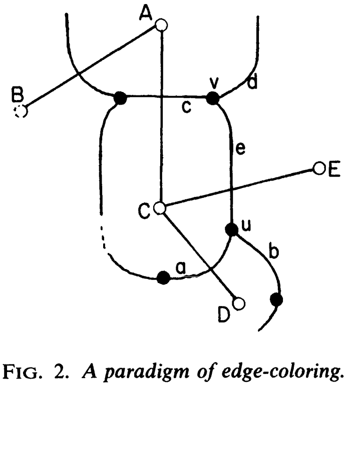


FIG. 2. *A paradigm of edge-coloring.*

case of a triangular face E with end-vertices u and v of the base e, u and v are incident each with $\Delta(G)-1$ edges already assigned colors, the same for both u and v. Fortunately, this cannot happen in a breadth-first traversal of $T(G)$ for an outerplane graph G with the maximum degree $\Delta(G) \geqq 5$.

LEMMA 1. *Let a 2-connected, outerplane graph G with the maximum degree $\Delta(G) \geqq 5$ be partially edge-colored by a breadth-first coloring algorithm using Δ colors. This Δ-coloring can be extended to a face E of G corresponding to the next-to-be visited node of $T(G)$, the associated tree of graph G.*

Proof. We proceed by induction on the number of visited nodes of $T(G)$. If E is the first face to be colored, then at most $3 < \Delta(G)$ colors are needed. Therefore, let us assume that C is the father node of E and the corresponding faces of G share an edge e with end vertices u and v (cf. Fig. 2). At most one of these two vertices may be incident with $\Delta(G)-1$ previously colored edges: if v is in the face corresponding to some ancestor of the node C, then u may be in at most one colored face other than C, namely, that corresponding to the left brother of node E. Thus, the number of previously colored edges incident with u is at most $3 < \Delta(G)-1$ (edges a, b, and e in Fig. 2), and edges of E can be colored using only $\Delta(G)$ colors. □

The above lemma does not translate directly for the case $\Delta(G)=4$, because of the distinct possibility that the base edge of a partially colored triangle face is adjacent to four colored edges forcing the same colors on both of the triangle's sides (cf. face 8 in Fig. 1). The edge-coloring during the visit of the corresponding node's father must prevent an occurrence of this situation. The coloring process will have to preserve the following property.

Property P_4. A partial 4-edge-coloring of a 2-connected outerplanar graph G with $\Delta(G)=4$ has property P_4 if and only if G does not have a colored edge (u, v) shared by a partially colored face E such that three colors are used to color all edges incident with vertices u and v.

In the breadth-first edge-coloring algorithm, property P_4 can be endangered only in two situations when coloring edges of the face corresponding to the father C of the node E. The first one, in which C's right brother would also be his leftmost brother (in the circular orientation of the root's children), is eliminated by rooting the associated tree in a leaf node. The second situation can be reached only through the sequence of face coloring (tree traversal) illustrated in Fig. 2. The node visiting order is $A \cdots BC \cdots DE$. Coloring edges of C we have to consider two cases of C's left brother B, which can be either external (corresponding to the outer face of the graph) or internal (see Fig. 2). In the former case, there is a choice of two colors for the first (leftmost) edge of C. This guarantees that the next to the last (rightmost) edge of C (edge a in Fig. 2) can be colored so as to preserve property P_4 (with respect to face E), namely by assigning it a color different from those assigned to edges c and d of face A. In the latter case, when the color of the first edge of C is forced by the formerly colored edges of A and B, we have additionally to consider the length of C. If C is a triangle, then the property P_4 might not be preserved. However, this property is not necessary for maintaining the property P_4 for coloring of E since C's leftmost son node (D in Fig. 2) cannot be interior. If C has length greater than 3, then there is enough freedom in coloring its edges to preserve property P_4. Thus, we have the following Lemma.

LEMMA 2. *Property P_4 can be preserved in a partial coloring of an outerplane graph G with $k=\Delta(G)=4$ colors following the breadth-first order coloring algorithm.*

An immediate corollary gives the desired statement about edge-coloring such graphs.

COROLLARY 3. *Let a 2-connected outerplanar graph G with the maximum degree $\Delta(G)=4$ be partially Δ-edge-colored by a breadth-first coloring algorithm. The Δ-edge-coloring can be extended to a face C of G corresponding to the next-to-be-visited node of $T(G)$.*

When $\Delta(G)=3$, the property of a partial 3-edge coloring required to avoid forced situations can be obtained directly from P_4.

Property P_3. A partial 3-edge-coloring of a 2-connected, outerplane graph G with $\Delta(G)=3$ has property P_3 if and only if G does not have a colored edge e shared by a not yet colored face such that the two colored edges adjacent to e have the same color.

LEMMA 4. *Let G' be a partially 3-edge-colored subgraph of a 2-connected, outerplane graph G with $\Delta(G)=3$. Let this partial 3-edge-coloring have property P_3 and be obtained through the breadth-first coloring algorithm. Let C be the next-to-be-visited node of $T(G)$. The 3-edge-coloring of G can be extended to C preserving property P_3.*

Proof. First, let us assume that C is the first face of G to be colored and that it has n edges. Let e be an edge shared by C and another face, C'. Such an edge always exists since $\Delta(G)=3>2$. We observe that the edges of C adjacent to e (or any other edge shared by another face) do not belong to any other face, because of the degree constraint. We assign color 2 to e and color other edges of C (say, clockwise) depending on the value of n. If $n \equiv 0 \bmod 3$, then coloring edges by 1–2–3 (with e appropriately included in the sequence) ensures property P_3. If $n \equiv 1 \bmod 3$, then we color edges of C by 1–2–3 starting with an edge adjacent to the initially colored e, but excluding e. The only two edges that could violate property P_3 are adjacent to e and thus belong only to C, since $\Delta(G)<4$. If $n \equiv 2 \bmod 3$, we again color edges along C by 1–2–3 starting with an edge adjacent to e. This time, however, the three last edges of C (i.e. f, g, h in Fig. 3) are colored differently, depending on adjacencies of f. If f is shared with another face, then it is colored 3, with colors 2 and 3 assigned to the remaining edges. Otherwise, the three edges are colored 2–1–3, respectively. Beside the edges adjacent to e, the only possibly offensive edges in the former case are adjacent to f and thus in no other face than C. In the latter case, only f and one edge adjacent to e have adjacent edges assigned the same color. By the remark above, they are in no other face and thus the property P_3 holds. Next, consider a G with property P_3 and a face C corresponding to the next-to-be-visited node of $T(G)$. Our inductive assumption yields that the three colored edges of C are assigned colors 3–2–1. The same case analysis as the one above proves that C can be colored to ensure property P_3. Thus G can be colored with three colors. □

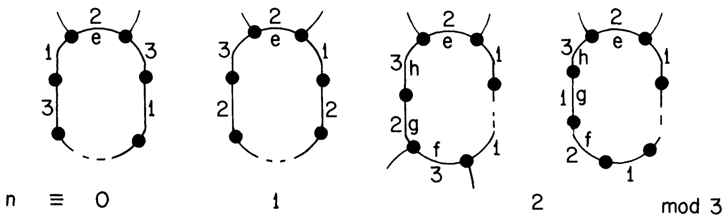


FIG. 3. *Coloring the first face of G.*

The amount of work necessary to color the edges of each face is proportional to the length of that face. Therefore, the total time spent edge-coloring an outerplanar graph is bounded by a linear function of the graph's size. Preprocessing a given outerplanar graph to obtain its rooted associated tree can also be performed in linear time (see Proskurowski and Sysło [11]). Collecting the results of this section, we finally have the following theorem.

THEOREM 2. *The breadth-first coloring algorithm produces an optimal edge-coloring of an outerplanar graph in time proportional to the size of the graph.*

REFERENCES

[1] K. APPEL AND W. HAKEN, *Every planar map is* 4-*colorable*, Illinois J. Math., 21 (1977), pp. 429-567.

[2] E. ARJOMANDI, *An efficient algorithm for colouring the edges of a graph with* $\Delta+1$ *colour*, INFOR, 20 (1982), pp. 82-101.

[3] S. FIORINI, *On the chromatic index of outerplanar graphs*, J. Combin. Theory B, 18 (1975), pp. 35-38.

[4] S. FIORINI AND R. J. WILSON, *Edge-Coloring of Graphs*, Research Notes in Mathematics 16, Pitman, London, 1977.

[5] H. N. GABOW AND O. KARIV, *Algorithms for edge coloring bipartite graphs and multigraphs*, SIAM J. Comput., 11 (1982), pp. 117-129.

[6] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability*, Freeman, San Francisco, 1979.

[7] F. HARARY, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.

[8] I. HOLYER, *The* NP-*completeness of edge-coloring*, SIAM J. Comput., 10 (1981), pp. 718-720.

[9] S. L. MITCHELL AND S. T. HEDETNIEMI, *Linear algorithms for edge-coloring trees and outerplanar graphs*, Inform. Proc. Lett., 9 (1979), pp. 110-112.

[10] T. NISHIZEKI, O. TERADA AND D. LEVEN, *Algorithm for edge-coloring graphs*, TRECIS 83001, Dept. Electrical Comm., Tokyo Univ., Tokyo, 1983.

[11] A. PROSKUROWSKI AND M. M. SYSŁO, *Minimum dominating cycles in outerplanar graphs*, Internat. J. Comput. Inform. Sci., 10 (1981), pp. 127-139.

[12] K. TAKAMIZAWA, T. NISHIZEKI, AND N. SAITO, *Linear-time computability of combinatorial problems on series-parallel graphs*, J. Assoc. Comput. Mach., 29 (1982), pp. 623-641.

[13] O. TERADA AND T. NISHIZEKI, *Approximate algorithms for the edge-coloring of graphs*, Trans. IECE Japan, J65-D, 11 (1982), pp. 1382-1389.

[14] A. WIDGERSON, *Improving the performance guarantee for approximate coloring*, J. Assoc. Comput. Mach., 30 (1983), pp. 729-735.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

PARALLEL ALGORITHMS FOR NONLINEAR PROBLEMS*

R. E. WHITE†

Abstract. Multi-splittings of a matrix are used to generate parallel algorithms to approximate the solutions of nonlinear algebraic systems. A parallel nonlinear Gauss-Seidel algorithm for approximating the solution of $Au+\phi(u)=f$ where A is an M-matrix is introduced and studied. Also, a parallel Newton-SOR method is defined for the problem $F(u)=0$ where $F'(u)=$ the Jacobian is an M-matrix. An illustration and comparison of these methods with their serial versions is given. The speed-up on the Denelcor HEP parallel processing computer is also recorded.

AMS(MOS) subject classifications. 65F10, 65H20

Key words. multisplitting, parallel nonlinear Gauss-Seidel algorithm, parallel Newton-SOR algorithm

1. Introduction. In this paper we discuss parallel algorithms (algorithms whose parts may be executed simultaneously) for the problems:

(1) $Au+\phi(u)=f$ where

A is an M-matrix,

$\phi(u)=(\phi_i(u_i))$ with $\phi_i:\mathbb{R}\to\mathbb{R}$ being continuous and nondecreasing,

$u,\ \phi(u),\ f\in\mathbb{R}^N$.

(2) $F(u)=0$ where

$F:\mathbb{R}^N\to\mathbb{R}^N$,

$F'(u)=(F_{iu_j}(u))$ is an M-matrix.

The importance of parallel algorithms is that on computers with parallel architectures convergence may be obtained in a shorter time than on serial computers. The main thrust of this paper is the development and study of two parallel algorithms which may be used to approximate the solutions to (1) and (2). The first algorithm (see (8)) generalizes the nonlinear Gauss-Seidel method. The second (see (16)) generalizes the Newton-SOR method. We briefly describe these two serial methods.

The *serial version of the nonlinear Gauss-Seidel* algorithm for the approximation of the solution of (1) is

(3) $u_i^{n+1}=r_i^{-1}(w)$ where

$$w=f_i-\sum_{j<i}a_{ij}u_j^{n+1}-\sum_{j>i}a_{ij}u_j^n,$$

$$r_i(z)\equiv a_{ii}z+\phi_i(z)=w,$$

$$z=r_i^{-1}(w)=\text{the inverse function of } r_i(z).$$

Under the conditions in (1) the iterative scheme in (3) converges to the unique solution of (1) (see J. M. Ortega and W. C. Rheinboldt [6] and [7]).

The *classical Newton method* for approximating the solution of (2) is

$$u^{n+1}=u^n-F'(u^n)^{-1}F(u^n), \tag{4}$$

* Received by the editors February 14, 1984, and in revised form September 15, 1984.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205.

or,

$$u^{n+1}=u^n-u^{n+1/2} \quad \text{where } F'(u^n)u^{n+1/2}=F(u^n).$$

If $F'(u)$ is an M-matrix, then $F'(u)=B(u)-C(u)$, where $B(u)\equiv D(u)-L(u)$ and $C(u)\equiv U(u)$ is the Gauss-Seidel splitting, is a regular splitting. In this case we have $\rho(B(u)^{-1}C(u))<1$, and hence

$$\begin{aligned} F'(u)^{-1}&=(B(u)-C(u))^{-1}\\ &=(B(u)(I-B(u)^{-1}C(u)))^{-1}\\ &=\sum_{m=0}^{\infty}(B(u)^{-1}C(u))^m B(u)^{-1}. \end{aligned}$$

Then we may approximate $F'(u)$ by truncating this series. This gives an approximation of $u^{n+1/2}$ and, consequently, generates the *serial version of the Newton*-SOR *method*

(5)

$$u^{n+1}=u^n-\sum_{m=0}^{M(n)-1}(B(u^n)^{-1}C(u^n))^m B(u^n)^{-1}F(u^n) \quad \text{where, for example, } M(n)=2^n.$$

In [6], [7] and A. Sherman [10] it was shown if $F'(u)=B(u)-C(u)$ is a weak regular splitting of the M-matrix and other conditions, then the algorithm (5) will converge. Furthermore, in [10] an error estimate was given (see (17) in Theorem 3).

There have been a number of papers dealing with parallel algorithms for linear problems. One of the first papers was by F. Robert [8] where a parallel algorithm based on a block iterative method was studied in the context of M-matrices. V. Conrad and Y. Wallach [2] have introduced an algorithm which was studied in terms of strict diagonal dominance. L. J. Hayes and P. Devloo [3] have recently used an overlapping block iterative method. This is similar to the schemes studied in D. P. O'Leary and R. E. White [5]. Some of this work is summarized in § 2 in this paper. Other parallel algorithms for linear problems are described in R. W. Hockney and C. R. Jesshope [4] and in A. Sameh [9].

In § 2 we review the results in O'Leary and White [5] on multi-splittings and parallel algorithms for linear problems. Section 3 contains the parallel nonlinear Gauss-Seidel algorithm, and § 4 has the parallel Newton-SOR algorithm. The last section contains an illustration and comparison of these algorithms to a problem which evolves from a semilinear elliptic boundary value problem.

2. Parallel linear algorithm. Let $S=\{1,\cdots,N\}$ correspond to the nodes of the algebraic problem $Au=f\in\mathbb{R}^N$. Suppose the nodes are grouped into K blocks where $S=$ the union of S_k with $k=1,\cdots,K$. These blocks may be overlapping and usually evolve naturally from the problem which generated the algebraic system. For example, if $Au=f$ comes from a partial differential equation on a rectangular grid, then the blocks may correspond to the rows or columns in the grid. Another example is from the finite element method where the blocks consist of the element nodes from groups of elements.

Often these blocks generate either a sequence of splittings of A

$$A=B_k-C_k, \qquad k=1,\cdots,K$$

or, a decomposition of A

$$A=\sum_{k=1}^{K}A_k.$$

In the latter case we may also generate a sequence of splittings of A by defining

$$B_k \equiv A_k + E_k$$

where E_k is a nonnegative diagonal matrix

$$C_k \equiv -\sum_{j \neq k} A_j + E_k.$$

If each B_k^{-1} exists, then we can define the algorithm

$$u^{k,n+1} \equiv B_k^{-1} C_k u^n + B_k^{-1} f \quad \text{with } u^n = u^{k,n}.$$

Because B_k emphasizes the components of A which are relevant to the nodes in S_k, this iterative scheme may converge very slowly. In order to accelerate the convergence, we introduce the weighting nonnegative diagonal matrices D_k where $\Sigma\, D_k = I$. (We will use the notation $\Sigma = \sum_{k=1}^{K}$.) Then we may weight each $u^{k,n+1}$ and compute the sum as

$$u^{n+1} \equiv \Sigma D_k u^{k,n+1}.$$

This leads us to state the following definitions.

DEFINITION. (B_k, C_k, D_k) is called a *multi-splitting* of A if and only if
i) $A = B_k - C_k$ where each B_k^{-1} exists,
ii) $\Sigma\, D_k = I$ where each $D_k =$ nonnegative diagonal matrix.

PARALLEL ALGORITHM. *Let (B_k, C_k, D_k) be a multi-splitting.*

$$\begin{aligned} u^{n+1} &\equiv (\Sigma\, D_k B_k^{-1} C_k) u^n + (\Sigma\, D_k B_k^{-1}) f \\ &= \mathbf{H} u^n + \mathbf{G} f \quad \text{where } \mathbf{H} \equiv \Sigma\, D_k B_k^{-1} C_k \text{ and } \mathbf{G} \equiv \Sigma\, D_k B_k^{-1}. \end{aligned} \tag{6}$$

Remarks. 1. We may use a SOR parameter to accelerate convergence

$$u^{n+1} = (1-w) u^n + w(\mathbf{H} u^n + \mathbf{G} f). \tag{6.1}$$

2. By using the properties of a multi-splitting we may write (6) as

$$u^{n+1} = u^n - \mathbf{G}(A u^n - f). \tag{6.2}$$

Or, if the SOR parameter is used, then (6.1) becomes

$$u^{n+1} = u^n - w\mathbf{G}(A u^n - f). \tag{6.3}$$

3. The terms in algorithm (6) may be computed simultaneously; hence, we call this a parallel algorithm.

In order to obtain convergence of (6), we must have $\rho(\mathbf{H}) < 1$. This is, in general, not true (see the example in [5]). Three different types of conditions (M-matrix, symmetric positive definite, and norm) each will essentially imply convergence of (6). These conditions are precisely stated in [5], but we restrict this paper to the M-matrix condition. Recall that an M-matrix $A = (a_{ij})$ is defined by the properties $a_{ij} \leqq 0$ for $i \neq j$ and $A^{-1} \geqq 0$. Also, $A = B - C$ is called a weak regular splitting of A when $B^{-1} \geqq 0$ and $B^{-1} C \geqq 0$. These two definitions are important because any weak regular splitting of an M-matrix implies $\rho(B^{-1} C) < 1$.

THEOREM 1. *If A is an M-matrix and $A = B_k - C_k$ $k = 1, \cdots, K$ are weak regular splittings of A, then $\rho(\mathbf{H}) < 1$.*

Examples. 1. Let $K=2$ and $N=2$.

$$A=\begin{pmatrix}2 & -1\\ -1 & 2\end{pmatrix}=\begin{pmatrix}2 & 0\\ -1 & 2\end{pmatrix}-\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}=B_1-C_1$$

$$=\begin{pmatrix}2 & -1\\ 0 & 2\end{pmatrix}-\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix}=B_2-C_2.$$

$$B_1^{-1}C_1=\begin{pmatrix}0 & \frac{1}{2}\\ 0 & \frac{1}{4}\end{pmatrix} \quad\text{and}\quad B_2^{-1}C_2=\begin{pmatrix}\frac{1}{4} & 0\\ \frac{1}{2} & 0\end{pmatrix}.$$

When

$$D_1=\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix} \quad\text{and}\quad D_2=\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix},$$

then

$$\mathbf{H}=\begin{pmatrix}0 & \frac{1}{2}\\ \frac{1}{2} & 0\end{pmatrix} \quad\text{and}\quad \rho(\mathbf{H})=\tfrac{1}{2}.$$

When

$$D_1=\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix} \quad\text{and}\quad D_2=\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix},$$

then

$$\mathbf{H}=\begin{pmatrix}\frac{1}{4} & 0\\ 0 & \frac{1}{4}\end{pmatrix} \quad\text{and}\quad \rho(\mathbf{H})=\tfrac{1}{4}.$$

When

$$D_1=\begin{pmatrix}\frac{1}{2} & 0\\ 0 & \frac{1}{2}\end{pmatrix}=D_2,$$

then

$$\mathbf{H}=\begin{pmatrix}\frac{1}{8} & \frac{1}{4}\\ \frac{1}{4} & \frac{1}{8}\end{pmatrix} \quad\text{and}\quad \rho(\mathbf{H})=\tfrac{3}{8}.$$

2. Let $A=D-L-U$ where D is a diagonal matrix, L is the strictly lower matrix and U is the strictly upper matrix. Let $A=(a_{ij})$ be an M-matrix and S_k be blocks of nodes where $k=1,\cdots,K$. Define the strictly lower submatrix

$$L_k\equiv\begin{cases}-a_{ij}, & j<i \text{ and } i,j\in S_k,\\ 0, & \text{otherwise.}\end{cases}$$

Let $A=(D-L_k)-(U+L-L_k)=B_k-C_k$ (see Fig. 1). Clearly, these are weak regular splittings of A. Also, if $K=1$ and $L_k=L$, then this is the Gauss-Seidel splitting; if $K=1$ and $L_k=0$, then this is the Jacobi splitting. Further, we restrict D_k:

Let d_i^k be the ith diagonal component of D_k.

Require $d_i^k=0$ if $i\notin S_k$.

Notation. Let b_{ij} be real numbers,

$$\Sigma^0 b_{ij}=\text{the sum with respect to } j\neq i \text{ and } j\notin S_k,$$

$$\Sigma^- b_{ij}=\text{the sum with respect to } j<i \text{ and } j\in S_k,$$

$$\Sigma^+ b_{ij}=\text{the sum with respect to } j>i \text{ and } j\in S_k,$$

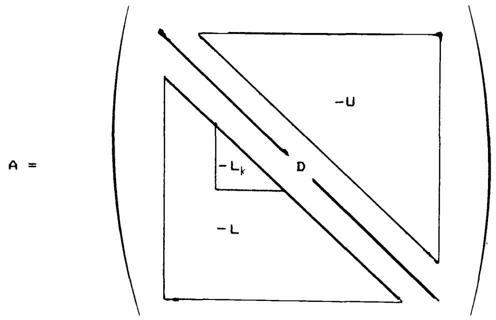


FIG. 1. *D, L, U and L_k.*

Then the parallel algorithm (6) may be written in component form as

$$u_i^{k,n+1} = [f_i - \Sigma^0 a_{ij} u_j^n - \Sigma^- a_{ij} u_j^{k,n+1} - \Sigma^+ a_{ij} u_j^n]/a_{ii}, \qquad (7)$$
$$u_i^{n+1} = \Sigma d_i^k u_i^{k,n+1}.$$

This will be generalized to the parallel nonlinear Gauss-Seidel algorithm which will be discussed in the next section.

3. Let $A = \Sigma A_k$ and $B_k \equiv A_k + E_k$ where E_k is a diagonal matrix with components e_i^k. When $C_k \equiv B_k - A$, then we have a sequence of splittings of A. This type of decomposition arises when A_k reflect rows or columns of grid points for a domain of a partial differential equation. The diagonal matrix is added to A_k so that each B_k^{-1} exists. Under the assumptions

i) $0 \leqq -a_{ij}^k \leqq -a_{ij}$, $i \neq j$ and $A_k = (a_{ij}^k)$,
ii) $e_i^k + a_{ii}^k > -\sum_{j \neq i} a_{ij}^k$,
iii) $e_i^k + a_{ii}^k > a_{ii}$,

it was shown in [5] that for any weighting matrices D_k the conditions of Theorem 1 hold. The essential properties of this type of decomposition are given in the following definition, and it will be used in the discussion of the parallel Newton-SOR algorithm in § 4.

DEFINITION. Let A, A_k, $E_k$and D_k be square matrices. (A_k, E_k, D_k) will be called a *convergent dissolution of* A if and only if

i) $A = \Sigma A_k$,
ii) E_k, $D_k \geqq 0$ are diagonal matrices,
iii) (B_k, C_k, D_k) is a multi-splitting of A where

$$B_k \equiv A_k + E_k \text{ and } C_k \equiv E_k - \sum_{j \neq k} A_j,$$

iv) $\rho(\mathbf{H}) < 1$ where $\mathbf{H} \equiv \Sigma D_k B_k^{-1} C_k$.

Remark. If $K = 1$, then any weak regular splitting of an M-matrix will be a convergent dissolution. The reader should consult A. Berman and R. J. Plemmons [1] as a general reference on M-matrices and splittings.

3. A parallel nonlinear Gauss–Seidel algorithm. In this section we consider the nonlinear algebraic problem (1). We assume A is an M-matrix, $\phi_i(u)$ are continuous and nondecreasing. When $\phi(u) = 0$, then we may apply the special parallel algorithm (7). If $\phi(u) \neq 0$, then we could apply the serial algorithm (3). The following algorithm is a combination of (3) and (7). We shall use the same notations as in the description of algorithm (7).

PARALLEL NONLINEAR GAUSS-SEIDEL ALGORITHM. Let $S=\{1,\cdots,N\}=$ the union of S_k. Let $r_i(z)\equiv a_{ii}z+\phi_i(z)=w$ and $r_i^{-1}(w)=z$. Assume $\Sigma d_i^k=1$, $d_i^k\geqq 0$ and $d_i^k=0$ for $i\notin S^k$. Let $i\in S_k$.

$$\begin{aligned} &u_i^{k,n+1}\equiv r_i^{-1}(w_i) \quad \text{where } w_i=f_i-\Sigma^0 a_{ij}u_j^n-\Sigma^- a_{ij}u_j^{k,n+1}-\Sigma^+ a_{ij}u_j^n,\\ &u_i^{n+1}\equiv \Sigma d_i^k u_i^{k,n+1}. \end{aligned} \tag{8}$$

An SOR parameter may be used as follows

$$u_i^{n+1}\equiv(1-w)u_i^n+w\Sigma d_i^k u_i^{k,n+1}. \tag{8.1}$$

THEOREM 2. *Let A be an M-matrix and $\phi_i:\mathbb{R}\rightarrow\mathbb{R}$ be continuous and nondecreasing. Then algorithm* (8) *converges to the unique solution of* (1).

Proof. By [7, Thm. 13.1.5], problem (1) has a unique solution and the serial algorithm (3) converges to the solution. Let $0\leqq L_k\leqq L$ be the lower triangular matrix associated with the block S_k (see § 2, Ex. 2, in this paper). Then the M-matrix A has the multi-splitting

$$A=(D-L_k)-(U+L-L_k)=B_k-C_k,$$

and for each k this is a regular splitting of A. Since A is an M-matrix, the assumptions of Theorem 1 hold, and hence $\rho(\mathbf{H})<1$.

The argument that follows is similar to the serial case in [7]. Since $a_{ii}>0$ and $\phi_i(z)$ are continuous and nondecreasing, then $r_i(z)$ are continuous and strictly increasing. Also, the following important inequalities are valid

$$\begin{aligned} a_{ii}|z-z^\sim|&\leqq|r_i(z)-r_i(z^\sim)|,\\ a_{ii}|r_i^{-1}(w)-r_i^{-1}(w^\sim)|&\leqq|w-w^\sim|. \end{aligned} \tag{9}$$

We use the notation for $f\in\mathbb{R}^N$ that $f=(f_i)$ where $f_i=[f]_i=$ the ith component of f. Then (8) may be written as

$$u_i^{k,n+1}=r_i^{-1}(w_i) \quad \text{where } w_i=[f+(U+L-L_k)u^n+L_k u^{k,n+1}]_i. \tag{10}$$

Let u be the solution of (1) and use the fact that (B_k, C_k, D_k) is a multi-splitting to obtain

$$u_i=r_i^{-1}(w_i^\sim) \quad \text{where } w_i^\sim=[f+(L+U-L_k)u+L_k u]_i. \tag{11}$$

Then lines (9), (10) and (11) combine to give

$$\begin{aligned} a_{ii}|u_i^{k,n+1}-u_i|&\leqq|[(L+U-L_k)(u^n-u)+L_k(u^{k,n+1}-u)]_i|\\ &\leqq[(U+L-L_k)|u^n-u|]_i+[L_k|u^{k,n+1}-u|]_i \end{aligned} \tag{12}$$

where $U+L-L_k$, $L_k\geqq 0$. Or, (12) in matrix form is, for $i\in S_k$,

$$[(D-L_k)|u^{k,n+1}-u|]_i\leqq[(U+L-L_k)|u^n-u|]_i. \tag{13}$$

Let $m\notin S_k$ and note that the mth row of $D-L_k$ has only one nonzero component, namely, the diagonal component. Consequently, (13) may be written as, for $i\in S_k$,

$$|u_i^{k,n+1}-u_i|\leqq[(D-L_k)^{-1}(U+L-L_k)|u^n-u|]_i. \tag{14}$$

Since for $i\in S_k$ and $\Sigma d_i^k=1$, we have from lines (8) and (14)

$$\begin{aligned} |u_i^{n+1}-u_i|&=\Sigma d_i^k|u_i^{k,n+1}-u_i|\\ &\leqq[\Sigma D_k B_k^{-1}C_k|u^n-u|]_i\\ &=[\mathbf{H}|u^n-u|]_i. \end{aligned} \tag{15}$$

Note (15) is independent of k, and by mathematical induction we have

$$\begin{aligned}|u^{n+1}-u| &\leqq \mathbf{H}|u^n-u| \\ &\ \vdots \\ &\leqq \mathbf{H}^n|u^1-u|.\end{aligned}$$

Since $\mathbf{H}\geqq 0$ and $\rho(\mathbf{H})<1$, algorithm (8) converges to the solution of (1).

4. Parallel Newton–SOR method. Let $F:\mathbb{R}^N\to\mathbb{R}^N$ and $F_i:\mathbb{R}^N\to\mathbb{R}$ for $i=1,\cdots,N$ be the component functions of F. Let $F'(u)=(F_{iu_j}(u))$ be the $N\times N$ Jacobian, the derivative matrix, of F. Consider problem (2) and suppose $F(u)=\Sigma F^k(u)=(\Sigma F_i^k(u))$ is a decomposition of $F(u)$. For example, if $F(u)=Au-f$ and $A=\Sigma A^k$, then define $F^1(u)=A^1u-f$, and $F^k(u)=A^ku$ for $k=2,\cdots,K$. In this case algorithm (6) has the form

$$\begin{aligned}u^{n+1}&=u^n-w\mathbf{G}_nF(u^n)\quad\text{where}\\ \mathbf{G}_n&\equiv\Sigma D_kB_k(u^n)^{-1}\quad\text{and}\\ B_k(u^n)&\equiv(F^k(u^n)+E_ku^n)'.\end{aligned}$$

Remarks. 1. If $K=1$, $w=1.0$ and $E_k=0$, then $B_1(u^n)=F'(u^n)$ and we retrieve Newton's method.

2. $B_k(u^n)^{-1}$ can be viewed as a crude approximation of

$$\begin{aligned}F'(u^n)^{-1}&=(B_k(u^n)(I-B_k(u^n)^{-1}C_k(u^n)))^{-1}\\ &=\left(\sum_{m=0}(B_k(u^n)^{-1}C_k(u^n))^m\right)B_k(u^n)^{-1}.\end{aligned}$$

If the series is truncated after $m=0$, then $F'(u^n)^{-1}\simeq B_k(u^n)^{-1}$.

Definition. Consider the nonlinear problem (2). A *parallel Newton*-SOR algorithm for approximating the solution is given by (16). Let $F(u)=\Sigma F^k(u)$ and let $F'(u)=B_k(u)-C_k(u)$ for $k=1,\cdots,K$ be a weak regular splitting of A where $B_k(u)$ may be defined as $(F^k(u)+E_ku)'$. Let $0\leqq m\leqq M(m)-1$, e.g. $M(m)=1$ or m or 2^m,

$$\begin{aligned}B_{k,n}&\equiv B_k(u^n),\\ C_{k,n}&\equiv C_k(u^n),\\ f_n&\equiv F(u^n)-(B_{k,n}-C_{k,n})u^n=F(u^n)-F'(u^n)u^n,\\ F_n(u)&\equiv(B_{k,n}-C_{k,n})u+f_n=F(u^n)+F'(u^n)(u-u^n),\\ u^{n+1,0}&\equiv u^n,\\ u^{n+1,m+1}&\equiv u^{n+1,m}-w\Sigma D_kB_{k,n}^{-1}F_n(u^{n+1,m}),\\ u^{n+1}&\equiv u^{n+1,M(m)}\quad\text{where } 0\leqq m\leqq M(n)-1 \text{ and } 1\leqq w<2.\end{aligned}\tag{16}$$

Notation. The iteration in (16) with respect to m will be called the *inner* iteration. The iteration with respect to n will be called the *outer* iteration. The *total number of inner* iterations for n outer iterations is $\sum_{m=0}^{n-1}M(m)$. In particular, when $M(m)=2^m$, then the total number of inner iterations is 2^n-1.

Remarks. 1. If $M(n)=1$ and $F(u)=Au-f$, then (16) is (6).

2. If $M(n)>0$, then we are just using a higher order approximation of $F'(u^n)^{-1}$.

3. If $M(n)$ is large, then there will be a smaller number of function evaluations.

4. If $F_n(u^{n+1,m})=0$, then $u^{n+1,m}=u^n-F'(u^n)^{-1}F(u^n)$ is just the next Newton iteration. Since this is the "best" one can hope for, the inner iteration should be terminated.

Theorem 3 establishes the convergence of algorithm (16) and gives estimates on the rate of convergence. In particular, line (17) implies superlinear convergence when $M(n)\to\infty$. The proof of Theorem 3 is nearly identical to the proof given in A. Sherman [10] where $K=1$. The asumptions used in this paper are similar, and assumptions 1, 2 and 3 are identical to Sherman's first three assumptions. The choice of norm and constants $r_0, \cdots, r_5$ will be explained in the sketch of the proof.

Assumptions. Let $F: R\to\mathbb{R}^N$ with $R\subset\mathbb{R}^N$ bounded. Let $u^*\in R$ and $F(u^*)=0$.

1. F is differentiable on $S_0\equiv\{u: \|u-u^*\|<r_0\}$.
2. $F'(u)$ is nonsingular at $u=u^*$.
3. There exists an $L>0$ such that for $u\in S_0$

$$\|F'(u)-F'(u^*)\|\leqq L\|u-u^*\|.$$

4. $F'(u)=\Sigma A_k(u)=B_k(u)-C_k(u)$ where

$$B_k(u)\equiv A_k(u)+E_k \quad\text{and}\quad C_k(u)\equiv-\sum_{j\neq k} A_j(u)+E_k.$$

5. Assume for $k=1,\cdots,K$ $B_k(u)$ are nonsingular at $u=u^*$.
6. $(A_k(u^*), E_k, D_k)$ is a convergent dissolution of $F'(u^*)$.
7. For $k=1,\cdots,K$ there exist $L_k>0$ and $r_5>0$ such that for

$$u\in S_5\equiv\{u: \|u-u^*\|<r_5\}$$

we have

$$\|B_k(u)-B_k(u^*)\|\leqq L_k\|u-u^*\|.$$

THEOREM 3. *Let assumptions* 1–7 *hold and* $F(u^*)=0$. *Let* $M(n)$ *be positive integers and define*

$$m^{\sim}=\max\left[\{1, M(0)\}\cup\left\{M(n)-\sum_{l=0}^{n-1} M(l): n=1,2,\cdots\right\}\right].$$

If $m^{\sim}<\infty$, *then there exist an* $r>0$ *and* $c<1$ *such that for* $u^0\in S\equiv\{u-u^*\|<r\}$, $w=1.0$ *and* u^{n+1} *given by* (16) *we have*

$$\|u^{n+1}-u^*\|\leqq c^{M(n)}\|u^n-u^*\|. \tag{17}$$

Sketch of the proof. We simply show that Theorem 3 falls into the context of Sherman's theorem. The first step is to show that algorithm (16) has the form given in Sherman's paper where $H(x)$ is replaced by $\mathbf{H}_n\equiv\Sigma D_kB_{k,n}^{-1}C_{k,n}$ and $B(x)^{-1}$ is replaced by $\mathbf{G}_n\equiv\Sigma D_kB_{k,n}^{-1}$. Consider (16) with $0\leqq m\leqq M(n)$.

$$\begin{aligned}
u^{n+1,m}&=u^{n+1,m-1}-\mathbf{G}_n((B_{k,n}-C_{k,n})u^{n+1,m-1}+f_n)\\
&=u^{n+1,m-1}-\Sigma D_k(u^{n+1,m-1}-H_{k,n}u^{n+1,m-1}+B_k^{-1}f_n), \qquad H_{k,n}\equiv B_{k,n}^{-1}C_{k,n}\\
&=\mathbf{H}_nu^{n+1,m-1}-\mathbf{G}_nf_n\\
&=\mathbf{H}_n(\mathbf{H}_nu^{n+1,m-2}-\mathbf{G}_nf_n)-\mathbf{G}_nf_n\\
&\ \ \vdots\\
&=\mathbf{H}_n^mu^n-\sum_{l=0}^{m-1}\mathbf{H}_n^l\mathbf{G}_nf_n.
\end{aligned}$$

Let $m=M(n)$.

$$\begin{aligned}
u^{n+1,M(n)} &= \mathbf{H}_n^{M(n)}u^n - \sum_{l=0}^{M(n)-1} \mathbf{H}_n^l \mathbf{G}_n f_n \\
&= u^n - \sum_{l=0}^{M(n)-1} \mathbf{H}_n^l[(I-\mathbf{H}_n)u^n + \mathbf{G}_n f_n] \\
&= u^n - \sum_{l=0}^{M(n)-1} \mathbf{H}_n^l[\Sigma D_k(I - B_{k,n}^{-1}C_{k,n})u^n \\
&\qquad + \Sigma D_k B_{k,n}^{-1}(F(u^n) - (B_{k,n} - C_{k,n})u^n)] \\
&= u^n - \sum_{l=0}^{M(n)-1} \mathbf{H}_n^l[\Sigma D_k B_{k,n}^{-1} F(u^n)] \\
&= u^n - \sum_{l=0}^{M(n)-1} \mathbf{H}_n^l \mathbf{G}_n F(u^n). \qquad (18)
\end{aligned}$$

Line (18) has the same form as line (2.6) in A. Sherman [10].

The second step is to discuss assumptions 4, 5, 6 and 7. The choice of r_5 in assumption 7 is to be smaller than $r_0, \cdots, r_4$ which are now defined. (Their existence is referenced in [10].). $r_1 \leqq r_0$ is defined so that $F'(u)$ is continuous and nonsingular on $S_1 \equiv \{u: \|u-u^*\| < r_1\}$. $r_2 \leqq r_1$ is defined so that on $S_2 \equiv \{u: \|u-u^*\| < r_2\}$ Newton's method converges Q-quadratically. $r_3 \leqq r_2$ is defined so that each $B_k(u)$ is continuous and nonsingular on $S_3 \equiv \{u: \|u-u^*\| < r_3\}$. Since assumption 6 holds, we have $\rho(\mathbf{H}(u^*)) < 1$. Consequently, there is a norm $\|\cdot\|$ such that $\|\mathbf{H}(u^*)\| < 1$. Adjust $r_0, \cdots, r_3$ so that the above results hold for this new norm. By a theorem in Ortega and Rheinboldt [7, pp. 350-351] there is a $r_4 \leqq r_3$ such that for $u \in S_4 \equiv \{u: \|u-u^*\| < r_4\}$ and u^{n+1} given by (18), i.e. (16), converges to u^*.

The third step is to establish the error estimate in (17). A careful inspection of Sherman's proof yields that it remains only to show the following:

There exist an $L^* > 0$ and a suitable neighborhood of u^* such that for u in this neighborhood we have

$$\|\mathbf{H}(u) - \mathbf{H}(u^*)\| \leqq L^*\|u-u^*\|. \qquad (19)$$

Since Assumptions 3, 4 and 5 hold, $B_k(u)^{-1}C_k(u) = I - B_k(u)^{-1}F'(u)$ is continuous in a neighborhood of u^*. Hence, $\mathbf{H}(u) = \Sigma D_k B_k(u)^{-1}C_k(u)$ is continuous in a neighborhood of u^*. Assumptions 4 and 5 imply $\|B_k(u)^{-1}\|$ is uniformly bounded in some neighborhood of u^*. Assumptions 3 and 7 imply $C_k(u)$ is Lipschitz continuous in some neighborhood of u^*. Consequently, for u in some neighborhood of u^* there exists L_k^* such that

$$\|B_k(u)^{-1}C_k(u) - B_k(u^*)^{-1}C_k(u^*)\| \leqq L_k^*\|u-u^*\|.$$

Thus, (19) must hold for some L^* and u in some neighborhood of u^*. This completes the sketch of the proof.

5. Numerical examples. In order to illustrate and compare the above algorithms, we consider the algebraic problem which evolves from a semilinear elliptic partial differential equation. The following problem is discretized by using the finite difference method.

$$-(K_1u_x)_x-(K_2u_y)_y=-ge^u \quad \text{on } \Omega,$$

$$u=x^2+y^2 \quad \text{on } \partial\Omega \quad \text{where}$$

$$K_1=1+x^2+y^2,$$

$$K_2=1+e^x+e^y,$$

$$g=2(2+3x^2+y^2+e^x+(1+y)e^y)e^{-x^2-y^2}$$

chosen so that $u=x^2+y^2$ is the unique solution,

$$\Omega=(0,1)\times(0,1)-[\tfrac{2}{5},\tfrac{3}{5}]\times[\tfrac{2}{5},\tfrac{3}{5}] \qquad \text{(see Fig. 2)},$$

$\partial\Omega$ = boundary with two parts.

In the finite difference method a rectangular grid is used with $\Delta x=\Delta y=1/nd=h$ where nd = the number of nodes in each direction. Unless otherwise indicated all the computations are for $nd=9$. If $nd=9$, then the number of unknowns is $9^2-3^2=72$; if $nd=19$, then the number of unknowns is $19^2-5^2=336$. If $K=1$, then the algorithm must be serial ((3) or (4) or (5)). If $K=24$ (for $nd=9$), then the blocks are from the 12 rows and the 12 columns. In this case the numerical scheme is a variation on the ADI method. If $K=4$, then the blocks are similar to the block S_1 whose 21 nodes are given by the dots in Fig. 2. The weighting matrices are given by $d_i^k=1/NK$ where NK is the number of elements in the set $\{k\colon i\in S_k, k=1,\cdots,K\}$. In all the computations convergence was defined by the l_2 error $<h^2$.

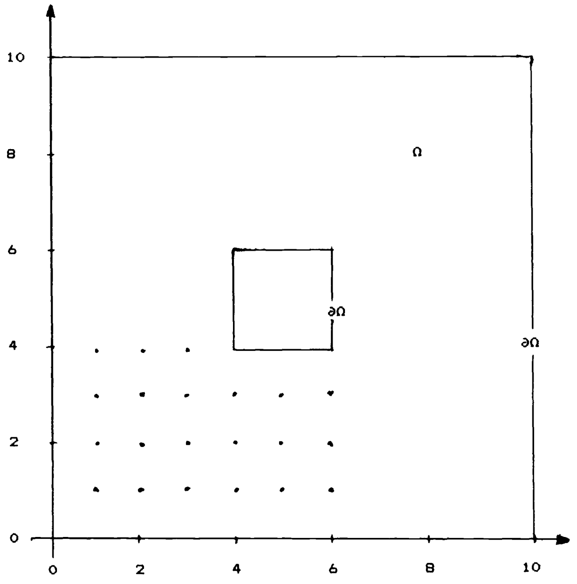


FIG. 2. *Ω with nodes of* S_1.

Table 1 indicates the number of iterations needed for convergence of the serial algorithms. The Newton and Newton-SOR computations were taken from A. Sherman [10]. In order to compare the serial and the parallel algorithms' computing times, one must consider (i) the operations for each serial iteration, (ii) the operations needed for each parallel iteration divided by the number of blocks, (iii) the number of iterations

TABLE 1
Serial algorithms

Algorithm	K	w	Number of iterations
Newton (3)	1	1.0	2
Gauss-Seidel (4)	1	1.2	33
Gauss-Seidel (4)	1	1.3	27
Newton-SOR (5)	1	1.4	3 outer (7 inner)

needed for convergence, and (iv) the overhead time for the parallel computer. Unfortunately, the latter significantly depends on the particular computer (see R. W. Hockney and C. R. Jesshope [4]). The results which are summarized in Tables 2-4 do not indicate the parallel overhead; they were computed on a serial computer. Table 5 contains some results obtained on the Denelcor HEP computer at Argonne National Laboratory.

Table 2 contains the iteration counts for the parallel nonlinear Gauss-Seidel algorithm (8). Note that the iteration count for $K=4$, $w=1.3$ is less than the serial nonlinear Gauss-Seidel algorithm (3) for $K=24$ and $w=1.3$.

TABLE 2
Parallel nonlinear Gauss-Seidel

K	w	Number of iterations
24	1.0	70
24	1.2	58
24	1.3	132
4	1.0	29
4	1.2	23
4	1.3	21
$4(nd=19)$	1.3	109

Table 3 lists the iterations needed for the parallel Newton-SOR algorithm (16) to converge. In both cases $M(m)=2^m$, and so the total number of inner iterations is 2^n-1 where $n=$the number of outer iterations.

TABLE 3
Parallel Newton-SOR

K	w	Number of iterations
24	1.3	4 outer (15 inner)
$48(nd=19)$	1.3	7 outer (127 inner)

Table 4 illustrates the superlinear convergence which is indicated by (17) in Theorem 3 for the parallel Newton-SOR method with $w=1.0$. In this table $nd=9$, $K=24$, $M(m)=2^m$ and $w=1.3$.

All the above results were simulated on a serial computer, and they only measure the convergence of the algorithm with no parallel overhead. The following numerical experiments were done on the Denelcor HEP multiprocessor at the Argonne National Laboratory. This particular machine consists of one process execution module (PEM) and simulates independent processors by pipelining instructions. The maximum

TABLE 4
Convergence rate for parallel Newton-SOR

n = outer iterations	l_2 error
1	3.90
2	1.64
3	2.97×10^{-1}
4	9.51×10^{-3}

effective use of this machine normally occurs when 8 to 12 processes are executing concurrently. We consider the same partial differential equation, but we changed the domain to $\Omega=(0,1)\times(0,1)$ (with no hole). This was done so that we could experiment with equal blocks where $K=1, 2, 4, 8, 16$. The convergence criteria was changed to the relative error being less than 0.0001 at each node. In order to measure the convergence of the algorithm and to measure the parallel overhead of the HEP, we introduce the following parameters.

Parallel Algorithm Index (PAI) $\equiv(I_1\times U_1)/(I_k\times U_k)$
I_k = iterations needed for convergence with k *equal* blocks
U_k = unknowns in each equal block.
Speed Up (SU) $\equiv E_0/E_k$
E_0 = execution time for serial algorithm with no parallel code
E_k = execution time for parallel algorithm with k blocks.

Since the blocks are equal, the products in PAI reflect the amount of work being done by each processor. When PAI is greater than 1.0, a savings in computing time by the parallel algorithm is indicated, provided the parallel overhead is ignored, i.e. the parallel overhead is assumed to be zero. The SU reflects both the convergence of the parallel algorithm and the parallel overhead. Thus, the difference PAI − SU may be viewed as a measure of the parallel overhead. For $K=1, 2, 4, 8, 16$ these values are tabulated in Table 5. The rapid increase of PAI − SU when K goes from 8 to 16 results from the HEP at Argonne only being able to simulate 8–12 independent processors.

TABLE 5
PAI *and* SU

K	Iter. for conv.	PAI	SU	PAI − SU
1*	57	1.0	1.0	0.0
1	57	1.0	0.999	0.001
2	58	1.769	1.617	0.152
4	63	2.931	2.337	0.594
8	69	4.461	2.915	1.546
16	70	7.329	2.919	4.410

* means that there is no parallel code

REFERENCES

[1] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[2] V. CONRAD AND Y. WALLACH, *Iterative solution of linear equations on a parallel processor*, IEEE Trans. Comp., C-26, (1977), pp. 838–847.

[3] L. J. HAYES AND P. DEVLOO, *An overlapping block iterative scheme for finite element methods*, Dept. Aerospace Engineering and Engineering Mechanics, Univ. Texas at Austin, to appear.

[4] R. W. HOCKNEY AND C. R. JESSHOPE, *Parallel Computers: Architecture, Programing, and Algorithms*, Adam Hilger Ltd., Bristol, 1981.

[5] D. P. O'LEARY AND R. E. WHITE, *Multi-splittings of matrices and parallel solution of liner systems*, this Journal, 6 (1985), pp. 630-640.

[6] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solutions on Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[7] ———, *Local and global convergence of general linear equations*, in Numerical Solution of Nonlinear Problems, J. M. Ortega and W. C. Rheinboldt, eds., Society for Industrial and Applied Mathematics, Philadelphia, 1970, pp. 122-143.

[8] F. ROBERT, *Methodes itératives « série parallèle »*, C. R. Acad. Sci., Paris, Ser. A 271 (1970), pp. 847-850.

[9] S. SAMEH, *An overview of parallel algorithms in numerical linear algebra*, in E. D. F. Bulletin de la direction des études et recherches, Ser. C 1 (1983), pp. 129-134.

[10] A. SHERMAN, *On Newton-iterative methods for the solution of systems of nonlinear equations*, SIAM J. Numer. Anal., 15 (1978), pp. 755-771.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

018

BALLOTING LABELLING AND PERSONNEL ASSIGNMENT*

W. D. WEI†, Y. Z. CAI‡, C. L. LIU§ AND A. M. ODLYZKO¶

Abstract. A personnel assignment problem is formulated as a problem of embedding a partially ordered set into another one. In this paper, an optimal solution to a special case in which the partially ordered sets are trees and forests is presented. Also, a related enumeration problem is studied.

Key words. combinatorial optimization, matching, balloting sequences

1. Introduction. The following problem was studied in Megiddo and Tamir [1]: $n=2m$ airplane pilots, linearly ordered by seniority, are to be divided into m teams each of which consists of a captain and a first officer. It is stipulated that the captain must have seniority over the first officer in each team. For each pilot, there is a measure of her effectiveness as a captain and a measure of her effectiveness as a first officer. We seek an assignment of the pilots that will maximize the total measure of effectiveness. Although the problem can be stated as a problem of maximum matching, it can also be formulated in a more general setting.

Let $(T, \leqq)$ be a set of personnel ordered by the partial ordering relation $\leqq$. Let $(P, \leqq)$ be a set of positions ordered by the partial ordering relation $\leqq$. We assume that $|T|=|P|$. An *assignment* ϕ is a one-to-one correspondence from T to P. An assignment ϕ is said to be *feasible* if $p_i \leqq p_j$ implies that $\phi^{-1}(p_i) \leqq \phi^{-1}(p_j)$. The positions in P are divided into r types, type-0, type-1, $\cdots$, and type-(r-1). Let h be a function from P to $\{0, 1, 2, \cdots, r-1\}$ such that $h(p_i)$ specifies the type of p_i. Let $f_0, f_1, \cdots, f_{r-1}$ be r functions from T to the reals such that $f_j(t_i)$ is a measure of t_i's effectiveness when t_i is assigned to a type-j position. Consequently, the total measure of effectiveness of an assignment is given by

$$\sum_{t_i \in T} f_{h(\phi(t_i))}(t_i). \tag{1}$$

In such a formulation, the problem of dividing pilots into m teams stated above is a special case in which $(T, \leqq)$ is a linearly ordered set where $\leqq$ is the linear ordering relation according to seniority, $(P, \leqq)$ is that shown in Fig. 1, and there are two types of positions corresponding to that of a captain and that of a first officer.

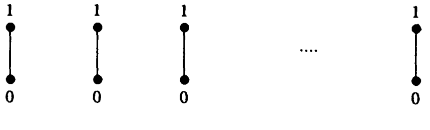


FIG. 1

The problem of determining a feasible assignment that maximizes the total measure of effectiveness is an NP-complete problem, as a matter of fact, even when $(T, \leqq)$ is

* Received by the editors January 18, 1983, and in revised form October 8, 1984.

† Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, and Department of Mathematics, Sichuan University, China.

‡ Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, and Wuhan Institute of Hydraulic and Electric Engineering, China.

§ Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801.

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

restricted to be a linearly ordered set and the value of r is restricted to 2 [2]. In [2], the case when $(T, \leqq)$ is a linearly ordered set and $(P, \leqq)$ belongs to a special class of partially ordered sets was studied. In this paper, we study a special case in which $(T, \leqq)$ is a tree.

2. An assignment problem. Let $(T, \leqq)$ be a rooted tree. Let $(P, \leqq)$ be as shown in Fig. 2. Namely, $(P, \leqq)$ is m copies of a rooted tree that has an internal node (the root) and k leaves. The positions corresponding to the roots are type-1 positions, and the positions corresponding to the leaves are type-0 positions. Let g be a function from T to the reals such that

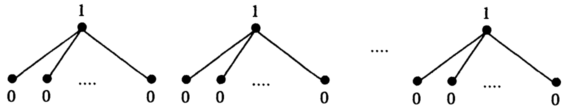


FIG. 2

$$g(t_i) = f_1(t_i) - f_0(t_i).$$

Then the expression in (1) can be rewritten as

$$\sum_{\substack{t_i \\ h(\phi(t_i))=1}} g(t_i) + \sum_{t_i} f_0(t_i).$$

Consequently, an optimal feasible assignment is one that maximizes the quantity

$$\sum_{\substack{t_i \\ h(\phi(t_i))=1}} g(t_i).$$

Given $(T, \leqq)$ and $(P, \leqq)$, $(T, \leqq)$ is said to be *feasible* if there exists at least one feasible assignment ϕ from T to P. Clearly, our optimization problem is meaningful only when $(T, \leqq)$ is feasible. We now proceed to show how to determine whether $(T, \leqq)$ is feasible, and if so, how to determine an optimal feasible assignment.

Given a tree $(T, \leqq)$ we assign to each node two numbers as follows:

(i) For a leaf t, $p(t) = -k$, $q(t) = 1$.

(ii) For an internal node t

$$p(t) = \sum_{v \text{ is a son of } t} q(v) - k,$$

$$q(t) = \begin{cases} p(t) & \text{if } p(t) \geqq 0, \\ p(t) + k + 1 & \text{if } p(t) < 0. \end{cases}$$

A node is said to be a *negative* node if $p(t) < 0$. A node is said to be a *nonnegative* node if $p(t) \geqq 0$.

LEMMA 1. *For any t, $q(t) \geqq 0$.*

Proof. The lemma is obviously true for a leaf t. Inductively, for an internal node t, since $q(v) \geqq 0$ for every son of t, $p(t) \geqq -k$. It follows that $q(t) \geqq 0$. □

Let $(T, \leqq)$ be a tree, and t be any node in T. We use T_t to denote the subtree of T rooted at t. Let $N(T_t)$ denote the total number of nodes in T_t and $N_1(T_t)$ denote the total number of negative nodes in T_t. We have

LEMMA 2. *For any node t in T,*

$$N_1(T_t) = \frac{1}{k+1}[k * N(T_t) + q(t)].$$

Proof. According to the definition,

$$q(t) = -k*(\text{number of leaves in } T_t)$$
$$-k*(\text{number of internal nodes in } T_t) + (k+1)*N_1(T_t).$$

That is,

$$(k+1)*N_1(T_t) = k*N(T_t) + q(t). \qquad \square$$

A (0, 1)-labelling of $(T, \leqq)$ is a mapping from T to $\{0, 1\}$. Let ϕ be an assignment from T to P. We use $h\phi$ to denote the (0, 1)-labelling of $(T, \leqq)$ such that $h\phi(t_i) = h(\phi(t_i))$. $h\phi$ will be referred to as the (0, 1)-labelling of T *induced* by ϕ. We have

LEMMA 3. *Let ϕ be a feasible assignment from T to P. Then, for the* (0, 1)*-labelling $h\phi$, in any subtree T_t the number of* 0's *in T_t is larger than or equal to k times the number of* 1's *in T_t.*

Proof. The lemma follows immediately from the fact that ϕ is feasible. □

LEMMA 4. *Let ϕ be a feasible assignment from T to P. Then, for the* (0, 1)*-labelling $h\phi$*

(i) *The number of* 0's *in T is equal to k times the number of* 1's *in T.*

(ii) *For any subtree T_t, the number of* 0's *in T_t is larger than or equal to the number of negative nodes in T_t.*

Proof. (i) is obvious.

(ii) can be proved by induction. If t is a leaf, clearly, (ii) is true. Inductively, let t be an internal node. If t is a nonnegative node, since (ii) is true for all subtrees rooted at the sons of t, (ii) is also true for T_t. If t is a negative node, let us examine all subtrees rooted at the sons of t. If in any one of these subtrees, (ii) is satisfied with strict inequality, that is, the number of 0's in that subtree is larger than the number of negative nodes in that subtree, then (ii) is true for T_t. On the other hand, consider the case in which in each subtree rooted at a son of t, the number of 0's in that subtree is equal to the number of negative nodes in the subtree. If $h(\phi(t)) = 1$, we have

$$\begin{aligned} \text{number of 0's in } T_t &= \sum_{v \text{ is a son of } t} \text{number of 0's in } T_v \\ &= \sum_{v \text{ is a son of } t} N_1(T_v) \\ &= \frac{1}{k+1}\left[k* \sum_{v \text{ is a son of } t} N(T_v) + \sum_{v \text{ is a son of } t} q(v)\right]. \end{aligned}$$

But $p(t) < 0$ implies that

$$\sum_{v \text{ is a son of } t} q(v) < k.$$

Thus,

$$\text{number of 0's in } T_t < \frac{1}{k+1}\left[k* \sum_{v \text{ is a son of } t} N(T_v) + k\right] = \frac{k}{k+1} N(T_t)$$

or

$$\text{number of 0's in } T_t < k*(\text{number of 1's in } T_t),$$

contradicting Lemma 3. Thus, we must have $h(\phi(t)) = 0$. In that case, (ii) follows immediately. □

We have now a criterion for $(T, \leqq)$ being feasible.

THEOREM 1. *$(T, \leqq)$ is feasible if and only if $q(t)=0$ where t is the root of T.*

Proof. Suppose $(T, \leqq)$ is feasible. Let $h\phi$ be a $(0,1)$-labelling induced by a feasible assignment ϕ. According to Lemma 2 and (ii) in Lemma 4,

$$\text{number of 0's in } T \geqq N_1(T) = \frac{1}{k+1}[k*N(T)+q(t)].$$

On the other hand, according to (i) in Lemma 4,

$$\text{number of 0's in } T = \frac{k}{k+1}N(T).$$

It follows that $q(t)=0$.

On the other hand, we shall show inductively that $(T, \leqq)$ is feasible if $q(t)=0$. Let $v_1, v_2, \cdots, v_i$ denote the sons of t. According to the definition,

$$q(v_1)+q(v_2)+\cdots+q(v_i)=k.$$

For $1 \leqq j \leqq i$, we show how we shall choose $q(v_j)$ nodes from T_{v_j}. We shall first remove from T_{v_j} any subtree rooted at a node with the value of its q-function being 0. Note that the removal of such subtrees does not alter the value of the q-function at each node. It follows that we can always find a path from v_j to a leaf in the resultant tree such that the value of the q-function at every node in the path is larger than 0. Removal of this leaf from the tree will reduce the value of the q-function of every node in the path by 1 and will not alter the value of the q-function of the other nodes in the tree. Such a step can be repeated $q(v_i)$ times until the value of the q-function at v_j becomes 0. Now, by the induction hypothesis, all subtrees rooted at nodes with the value of their q-functions being 0 are feasible. Furthermore, the root t of T, together with the $q(v_1)+q(v_2)+\cdots+q(v_i)=k$ nodes removed as described above, can be assigned to a rooted tree with k leaves. □

Let ρ be a $(0,1)$-labelling of $(T, \leqq)$. ρ is said to be a *balloting* labelling[1] of $(T, \leqq)$ if according to ρ

(i) The number of 0's in T is equal to k times the number of 1's in T.

(ii) For any node t, the number of 0's in T_t is larger than or equal to the number of negative nodes in T_t.

LEMMA 5. *Let ρ be a balloting labelling of $(T, \leqq)$. Then there is a feasible assignment ϕ from $(T, \leqq)$ to $(P, \leqq)$ such that $h\phi$ is equal to ρ.*

Proof. Let t be any vertex in T. In T_t, according to the labelling ρ,

$$\text{number of 0's in } T_t \geqq N_1(T_t) \geqq \frac{k}{k+1}N(T_t).$$

Thus

$$\text{number of 1's in } T_t \leqq N(T_t) - \frac{k}{k+1}N(T_t) = \frac{1}{k+1}N(T_t).$$

It follows that

$$\text{(2)} \qquad \text{number of 0's in } T_t \geqq k*(\text{number of 1's in } T_t).$$

We now show how to construct a feasible assignment from $(T, \leqq)$ to $(P, \leqq)$ by showing how each vertex t in T such that $\rho(t)=1$ can be matched with k vertices

[1] The reason for such a choice of terminology will become obvious later.

$t_1, t_2, \cdots, t_k$ such that $\rho(t_1)=\rho(t_2)=\cdots=\rho(t_k)=0$ and $t \leqq t_1, t \leqq t_2, \cdots, t \leqq t_k$. We use the following algorithm:

(i) Initially, all vertices in T are unmarked.

(ii) Let t be a vertex in T such that t is unmarked, $\rho(t)=1$, and T_t contains no other vertex t' such that t' is unmarked and $\rho(t')=1$.

(iii) According to (2), there are k or more unmarked vertices in T_t, $t_1, t_2, \cdots, t_k$ such that $\rho(t_1)=\rho(t_2)=\cdots=\rho(t_k)=0$. Match t with $t_1, t_2, \cdots, t_k$, and mark $t, t_1, t_2, \cdots, t_k$.

(iv) Repeat (ii) and (iii) until all vertices in T are marked. □

Lemmas 3 and 4 form the basis of an algorithm for determining an optimal feasible assignment from $(T, \leqq)$ to $(P, \leqq)$, since a balloting labelling of T, ρ, that maximizes the quantity

$$\sum_{T_i \in T} \rho(t_i) g(t_i)$$

will yield a (not necessarily unique) optimal feasible assignment. We use the following algorithm to determine such a balloting labelling ρ.

(i) If t is a leaf, set $\rho(t)=0$, and mark t.

(ii) Let t be a negative internal node such that in T_t there are no *unmarked* negative nodes. Among all nodes in T_t that have yet not been labelled by ρ, choose a vertex v such that

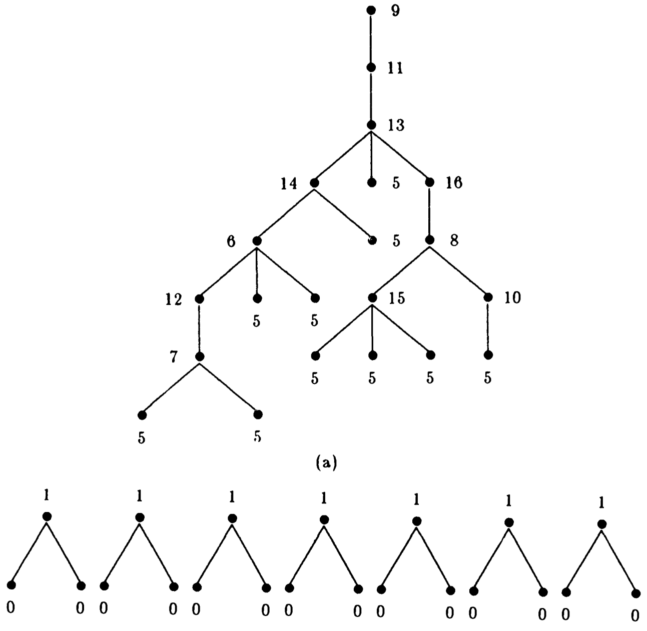
$$g(v) = \min_{\substack{t_i \in T_t \\ t_i \text{ is not labelled}}} g(t_i),$$

label $\rho(v)=0$, and mark t.

(iii) Repeat (ii) until all negative internal nodes are marked.

(iv) Label all remaining nodes with 1.

As an illustrative example, consider the tree $(T, \leqq)$ shown in Fig. 3a, where the number associated with each node t in T is the value of $g(t)$, and the forest $(P, \leqq)$

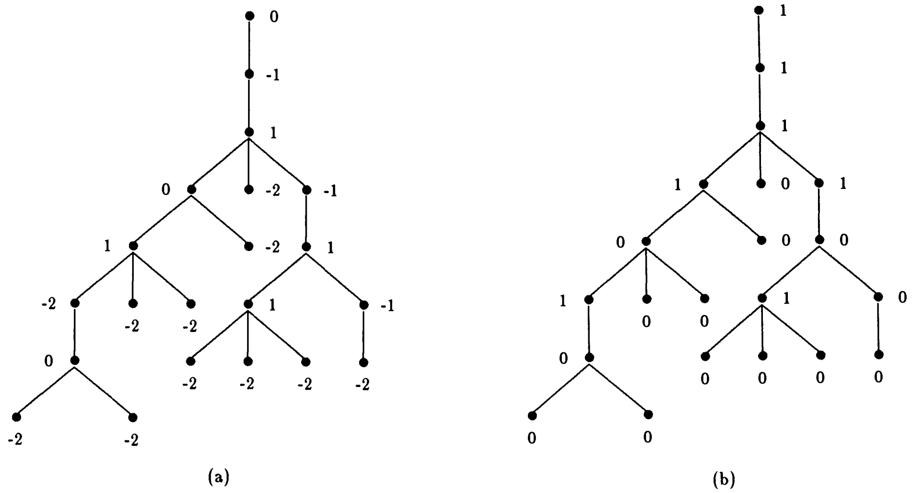


(a)

(b)

FIG. 3

shown in Fig. 3b, where the number associated with each node p in P is the type of position p. Fig. 4a shows the value of $p(t)$ at each node. Fig. 4b shows the $(0, 1)$-labelling $h\phi$ corresponding to an optimal feasible assignment ϕ.



(a)

(b)

FIG. 4

3. An enumeration problem. A balloting sequence is a sequence of equal numbers of 0's and 1's such that in any prefix of the sequence the number of 1's is larger than or equal to the number of 0's. Equivalently, in any suffix of a balloting sequence the number of 0's is larger than or equal to the number of 1's. We generalize the notion of a balloting sequence to that of a balloting labelling of a tree. In particular, we are interested in counting the number of balloting labellings of a rooted, regular, full k-ary tree T. The following lemma will make it more obvious that the notion of a balloting labelling is a natural extension of a balloting sequence.

LEMMA 6. *Let ρ be a $(0, 1)$-labelling of a tree $(T, \leqq)$. Then ρ is a balloting labelling if and only if*

(i) *The number of* 0*'s in T is equal to k times the number of* 1*'s in T.*

(ii) *For any node t, the number of* 0*'s in T_t is larger than or equal to k times the number of* 1*'s in T_t.*

Proof. The if part is almost identical to the proof of Lemma 4 (simply replace $h\phi$ by ρ). The only if part was included in the proof of Lemma 5. □

Thus, indeed, for the case $k = 1$, the number of balloting labellings becomes the famous Catalan number. A $(0, 1)$-labelling of a tree such that only the condition in (ii) of Lemma 6 is satisfied is referred to as an *unbalanced* balloting labelling.

We are not able to obtain a closed form expression for the number of balloting labellings of a rooted, regular, full k-ary tree. However, our development in § 2 enables us to carry out a recursive computation. We shall illustrate the computational procedure for the case of binary trees.

Let T_i denote the rooted, regular, full binary tree of height i. Let $s(n, m)$ denote the number of unbalanced balloting labellings of T_{2n+1} with $N_1(T_{2n+1}) + m$ 0's. Let $r(n, m)$ denote the number of unbalanced balloting labellings of T_{2n} with $N_1(T_{2n}) + m$ 0's. Note that $s(n, 0)$ is the number of balloting labellings of the rooted, regular, full binary tree of height $2n+1$. Also, note that

$$\begin{aligned} &s(0,0)=1,\\ &s(0,1)=1,\\ &s(0,i)=0 \quad \text{for } i\geqq 2, \end{aligned}$$

and

$$s(n,i)=0, \qquad r(n,i)=0 \quad \text{for } i<0.$$

LEMMA 7.

$$s(n,m)=\sum_{m_1+m_2=m-1} r(n,m_1)r(n,m_2)+\sum_{m_1+m_2=m} r(n,m_1)r(n,m_2),$$

$$r(n,m)=\sum_{m_1+m_2=m+1} s(n-1,m_1)s(n-1,m_2)+\sum_{m_1+m_2=m} s(n-1,m_1)s(n-1,m_2).$$

Proof. Note that

$$\begin{aligned} &N_1(T_{2n+1})=\tfrac{2}{3}(2^{2n+2}-1),\\ &N_1(T_{2n})=\tfrac{1}{3}(2^{2n+2}-1),\\ &N_1(T_{2n-1})=\tfrac{2}{3}(2^{2n}-1). \end{aligned}$$

The lemma follows from examining the two possibilities of labelling the root of T_{2n+1} (and T_{2n}) with a 0 and a 1, respectively. □

Lemma 7 enables us to compute $s(n,m)$ and $r(n,m)$ recursively. For example, we have computed:

$$\begin{aligned} &s(0,0)=1,\\ &s(1,0)=3^2,\\ &s(2,0)=(3^4*7)^2,\\ &s(3,0)=(3^{15}*7^3*331)^2,\\ &s(4,0)=(3^{60}*7^{11}*11*331^3*7417)^2,\\ &s(5,0)=(3^{238}*7^{44}*11^3*61*331^{11}*7417^3*312781459)^2, \qquad (3)\\ &s(6,0)=(3^{951}*7^{174}*11^{11}*61^3*331^{43}*7417^{11}\\ &\qquad *312781459^3*25953749*510438906725663)^2,\\ &s(7,0)=(3^{3801}*7^{694}*11^{43}*61^{11}*331^{171}*7417^{43}*312781459^{11}\\ &\qquad *25953749^3*510438906725663^3*5*103*563\\ &\qquad *121357460363579295946410138878592378911 77)^2,\\ &s(8,0)=7.57631795308851224*10^{26142}. \end{aligned}$$

Let

$$S_n(z)=\sum_{m\geqq 0} s(n,m)z^m,$$

$$R_n(z)=\sum_{m\geqq 0} r(n,m)z^m.$$

According to Lemma 7

$$(4) \qquad S_n(z) = zR_n^2(z) + R_n^2(z),$$

$$(5) \qquad R_n(z) = z^{-1}[S_{n-1}^2(z) - S_{n-1}^2(0)] + S_{n-1}^2(z).$$

Combining (4) and (5), we obtain

$$(6) \qquad S_n(z) = (z+1)[(1+z^{-1})S_{n-1}^2(z) - z^{-1}S_{n-1}^2(0)]^2.$$

The equation in (6) shows that

$$s(n,0) = s(n-1,0)^2[s(n-1,0) + 2s(n-1,1)]^2 \geqq s(n-1,0)^4,$$

and

$$s(n,0)^{4^{-n}} \geqq s(n-1,0)^{4^{-(n-1)}}.$$

Therefore, $s(n,0)^{4^{-n}}$ is an increasing function of n, and the limit

$$\lim_{n\to\infty} s(n,0)^{4^{-n}}$$

exists.

Using any one of the values of $s(n,0)$ in (3), we can obtain a lower bound of $s(n,0)$. For example, we have computed:

$$s(n,0) \geqq (2.496937)^{4^n} \quad \text{for } n \geqq 5,$$

$$s(n,0) \geqq (2.503241)^{4^n} \quad \text{for } n \geqq 6,$$

$$s(n,0) \geqq (2.505582)^{4^n} \quad \text{for } n \geqq 8.$$

On the other hand, we can also compute an upper bound of $s(n,0)$. From (6) we have

$$(7) \qquad S_n(z) \leqq \frac{(z+1)^3}{z^2} S_{n-1}(z)^4 \quad \text{for } z > 0.$$

Applying (7) repeatedly, we obtain

$$S_n(z) \leqq \left[\frac{(z+1)^3}{z^2}\right]^{1+4+4^2+\cdots+4^{n-n_0-1}} S_{n_0}(z)^{4^{n-n_0}}$$

$$= \left[\frac{z^{2/3}}{1+z}\right]^{1-4^{n-n_0}} S_{n_0}(z)^{4^{n-n_0}} = \frac{z^{2/3}}{(1+z)}\left[\frac{z+1}{z^{2/3}} S_{n_0}(z)\right]^{4^{n-n_0}},$$

that is

$$S_n(z) \leqq \left[\frac{z+1}{z^{2/3}} S_{n_0}(z)\right]^{4^{n-n_0}} \quad \text{for } n_0 \leqq n \text{ and } 0 < z < 1.$$

Since from (6) we have $s(n,0) \leqq S_n(z)$ for $z > 0$

$$s(n,0) \leqq \left[\left(\frac{z+1}{z^{2/3}} S_{n_0}(z)\right)^{4^{-n_0}}\right]^{4^n} \quad \text{for } n_0 \leqq n, \quad 0 < z < 1.$$

If $s(n,0)$ is known for $0 \leqq n \leqq n_0 - 1$, then (6) enables us to compute $S_{n_0}(z)$ for any given $z > 0$. For $n_0 = 6$ and $z = 3 * 10^{-3}$ we obtain

$$s(n,0) \leqq (2.505992)^{4^n}$$

for $n \geqq 6$. Using the numerical values in (3) for $s(n,0)$ for $n = 0-5$, we confirm that the bound is actually valid for all $n \geqq 0$.

For $n_0 = 9$ and $z = 2.5 * 10^{-4}$, we obtain a slightly tighter bound.

$$s(n, 0) \leqq (2.505786)^{4^n}, \qquad n \geqq 0.$$

4. Remarks. The problem studied in this paper can also be formulated as a matroid optimization problem. For a rooted tree $(T, \leqq)$ and any integer $k > 0$, let $\mathcal{F}$ denote the family of all subsets F of T such that for every node $t \in T$, $k \cdot |T_t \cap F| \leqq |T_t - F|$ where T_t is the set of all nodes in the subtree rooted at t. (Note that in the terminologies of § 2, each subset F corresponds to a (0, 1)-labelling of T in which the nodes in F are labelled 1.) It is not difficult to show that $\mathcal{F}$ is the family of independent sets of a matroid on T. Thus, our optimization problem becomes that of finding a maximum weight independent set of size $|T|/(k+1)$, which can be solved using the "greedy" algorithm for matroids. (We are grateful to a referee who pointed out to us such a formulation.)

It should also be noted that we solve in this paper only a special case of an assignment problem. The case in which $(P, \leqq)$ contains trees of height larger than 1, and the case in which there are more than 2 types of positions are unsolved. Such cases will probably require the development of new methods of solution other than that presented here.

REFERENCES

[1] N. MEGIDDO AND A. TAMIR, *An $O(n \log n)$ algorithm for a class of matching problems*, SIAM J. Comp., 7 (1978), pp. 154–157.

[2] P. RAMANAN, J. S. DEOGUN AND C. L. LIU, *A personnel assignment problem*, J. Algorithms, 5 (1984), pp. 132–144.

[3] C. L. LIU, *Introduction to Combinatorial Mathematics*, McGraw-Hill, New York, 1968.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 1, January 1986

019

GENERALIZED BINARY BINOMIAL GROUP TESTING*

NADER MEHRAVARI†

Abstract. The conventional group testing problem is that of correctly classifying each member of a given population as defective or non-defective. A conventional binary group test is a simultaneous test on a subset of the population with only two possible outcomes. A "good" reading indicates that all the members of the subset are non-defective, and a "bad" reading shows that there is at least one defective member in the subset. The goal is to design an efficient algorithm to correctly identify all the defective members of a population. In this paper, we introduce the idea of generalized binary binomial group testing. The generalized group tests provide different information about the number of defectives in a group than does the conventional group test. In particular, motivated by problems in finite-user random-access communication systems, we investigate the following two generalized binary group tests: the so-called conflict/no conflict test which indicates whether there is at most one defective item in a group, and the so-called success/failure test which indicates if there exists exactly one defective item in a group. We introduce and analyze group testing procedures for the above generalized group testing problems. The proposed procedures perform better than the scheme of testing each item individually and the algorithms based on binary tree search methods. Optimality of the proposed algorithms is also discussed.

Key words. group testing, blood testing, random-access communication

AMS(MOS) subject classifications: 62, 60, 5

1. Introduction. The problem of conventional group testing is concerned with correctly classifying each of the units in a population of size M as defective or non-defective. In the binomial group testing problem, each unit represents an independent Bernoulli trial with probabilities p and $q=1-p$ of being defective and non-defective, respectively. A conventional binary group test is a simultaneous test on n units with only two possible outcomes. A "good" reading indicates that all n units are non-defective, and a "bad" reading shows that there is at least one defective unit among the group. The goal is to design sequential testing procedures that minimize the expected number of group tests. Hereafter, the above problem will be referred to as conventional binary binomial group testing or simply conventional group testing.

Group testing was first introduced during World War II by Dorfman [1]. Dorfman introduced a method that identified all syphilitic men called up for induction using up to 80 percent fewer blood tests than in the previously-employed method of testing each individual. In Dorfman's scheme, after the blood samples were drawn, they were pooled in groups of n, whereupon groups, rather than individuals, were subjected to the test. If none of the n individuals in the group were syphilitic, then the test would be negative. If, however, one or more of the individuals in the group carried the syphilitic antigen, the test would be positive. In this latter case, the individuals in that group had to be tested individually. Dorfman computed the most efficient group size, n, and showed that on the average, his scheme required fewer blood tests.

Sterrett [4] improved upon Dorfman's procedure by testing individual members of a defective set only until a defective unit was found. (A defective set is one which is

*Received by the editors September 1, 1983, and in final revised form February 15, 1985. This paper was typeset at AT&T Bell Laboratories, Holmdel, New Jersey, using the *troff* program running under the UNIX® operating system. Final copy was produced on July 10, 1985.

†AT&T Bell Laboratories, Holmdel, New Jersey 07733.

known to contain at least one defective item.) Then the remaining units from that defective set were pooled and tested. This was continued until that particular defective set was completely analyzed. Sobel and Groll [3] further generalized this idea by testing small subsets of a defective set rather than immediately testing individual units.

In Section 2, we introduce the idea of the generalized binary group testing problem. In particular, we describe two generalized group tests that were motivated by problems in the area of random-access communication systems under a finite-user model [2]. The first is referred to as the Conflict/No Conflict (CNC) test, and it indicates whether there is at most one defective unit in a group. The second is referred to as the Success/Failure (SF) test, and it shows whether there is exactly one defective item in a group. Sections 3 and 4 contain the description and analysis of group testing procedures employing CNC and SF tests, respectively. The proposed procedures perform better than the method of testing individual items and the methods based on binary tree search for small values of the probability, p, that an item is defective, and adapt themselves to testing one item at a time for higher values of p. In addition, the optimality of these procedures is discussed.

2. Generalized binary binomial group testing problem. In this section, we introduce the idea of a generalized binary group test and formalize the corresponding binomial group testing problem. Hereafter, we make use of the following terminology:

1. A non-defective set is a set that contains no defectives.
2. A defective set is a set that contains at least one defective item among its members.
3. A conflicted set is a set that contains at least two defective items among its members.
4. A binomial set is a set whose members are each defective with probability p independently of one another.

DEFINITION 2.1. Let $R(G)$ be the random variable representing the number of defective items in a group G. A generalized binary group test $T(a,b;G)$ is a simultaneous test on group G with only two possible outcomes. A "good" reading indicates that $a \leqslant R(G) \leqslant b$, and a "bad" reading shows otherwise (i.e., $R(G) < a$ *or* $R(G) > b$), where the integers a and b satisfy $0 \leqslant a \leqslant b \leqslant |G|$.

By choosing the integer pair (a,b), the generalized test $T(a,b;G)$ will provide us with specific binary group tests which provide different kinds of information. Three of these tests that are the subject of this paper are identified below.

1. Something/Nothing (SN) Test: Let $a=0$ and $b=0$ in the definition of $T(a,b;G)$. The resulting test indicates whether there is at least one defective in a group. Note that the SN test coincides with the conventional group test.
2. Conflict/No Conflict (CNC) Test: Let $a=0$ and $b=1$ in the definition of $T(a,b;G)$. The resulting test indicates whether there are at least two defectives in a group.
3. Success/Failure (SF) Test: Let $a=1$ and $b=1$ in the definition of $T(a,b;G)$. The resulting test indicates whether there is exactly one defective in a group.

The SN group testing problem that coincides with the conventional group testing problem has been studied by many authors including Dorfman [1], Sterrett [4], Sobel and Groll [3], and Ungar [5]. A simple procedure was proposed by Sobel and Groll [3] for solving the SN problem that has been shown to be optimum (in the sense of minimizing the expected number of group tests) for large values of p and to be close to the (yet unknown) optimum strategy for small values of p. Let $H_{SN}(M)$ represent the expected number of group tests needed to classify a population of size M by using this procedure. For the sake of completeness and comparison, the values of $H_{SN}(M)$ for $M=64$ have been illustrated in Figure 1 as a function of p. This procedure is shown to be optimum for the range of $p \geqslant (3-\sqrt{5})/2$ where the strategy adapts itself to testing individual items [5].

3. Conflict/No Conflict group testing. Consider the CNC group test that was introduced in the previous section. This particular test was motivated by the following problem in the area of random-access communication systems under a finite-user model [2]. In such systems, simultaneous transmission by two or more active transmitters results in interference. The task is to partition a population of transmitters into subsets so that each subset contains at most one active transmitter. Then, granting transmission rights to one subset at a time would result in interference-free transmission. We now want to investigate the generalized group testing problem that was motivated by this communication problem. A CNC group testing problem is defined as follows:

DEFINITION 3.1. A Conflict/No Conflict (CNC) group testing problem is concerned with partitioning a population of size M into subsets such that each subset contains at most one defective object. Members of the population are each associated with an independent Bernoulli $(p,q=1-p)$ random variable, where p is the probability that an object is defective. We have at our disposal a Conflict/No Conflict generalized binary group test $T(0,1;G)$. The goal is to design efficient algorithms for correctly partitioning the population.

We will consider the problem only for finite M where efficiency is defined in the sense of minimizing the expected number of group tests needed to partition the population. CNC group testing differs from conventional (SN) group testing since the latter is capable of identifying the individual defective items, whereas the former is only capable of distinguishing between groups of items with at most one defective (non-conflicted group) and groups with two or more defectives (conflicted group). In addition to the application of CNC group testing in random-access communication systems, this type of group testing can be used in any industrial and/or medical setting where the presence of two or more items (devices, molecules, cells, insects, etc.) having the same property would be undesirable and needs to be avoided in the most efficient way.

We now propose and analyze an algorithm for the CNC group testing problem. The proposed procedure is in the same class as the one introduced by Sobel and Groll[3] for the conventional (SN) group testing problem. In what follows, we use the same definitions and terminology used by Sobel and Groll. The definitions of a defective set, a non-defective set, a conflicted set, and a binomial set are the same as those in Section 2. If a subset S of a conflicted set T is known to contain at most one defective, the set $T-S$ is called a "mystery" set. The proposed procedure has the property that, at each step, the unclassified portion of the population is divided into a binomial set, a conflicted set, and a mystery set. Let n and m represent the number of unclassified items and the size of the conflicted set at some step of the algorithm, respectively.

Let random variables Z and V represent the number of defectives in two disjoint binomial sets of size x and $m-x$, respectively. Let the random variable Y be equal to $Z+V$, i.e., Y represents the number of defectives in a binomial set of size m. To describe the proposed protocol, we introduce the symbols $\alpha(x)$, $\gamma(x,m)$, $\delta(x,m)$ which are defined as:

$$\alpha(x) = P(Z \geqslant 2),$$

$$\gamma(x,m) = P(Z \geqslant 2 | Y \geqslant 2),$$

$$\delta(x,m) = P(V \geqslant 2 | Y \geqslant 2,\ Z \leqslant 1).$$

These quantities can easily be expressed in terms of p,

$$\alpha(x) = 1 - q^x - xpq^{x-1}, \qquad x \geqslant 2, \tag{3.1}$$

$$\begin{aligned}\gamma(x,m) = P(Z \geqslant 2 | Y \geqslant 2) &= \frac{P(Z \geqslant 2,\ Y \geqslant 2)}{P(Y \geqslant 2)} \\ &= \frac{P(Z \geqslant 2,\ V \geqslant 0)}{P(Y \geqslant 2)} \\ &= \frac{P(Z \geqslant 2)}{P(Y \geqslant 2)} = \begin{cases} \dfrac{\alpha(x)}{\alpha(m)} & \text{for } x \geqslant 2, \\ 0 & \text{for } x = 1. \end{cases}\end{aligned} \tag{3.2}$$

$$\delta(x,m) = P(V \geqslant 2 | Y \geqslant 2,\ Z \leqslant 1) = \frac{P(V \geqslant 2,\ Y \geqslant 2,\ Z \leqslant 1)}{P(Y \geqslant 2,\ Z \leqslant 1)}$$

$$= \frac{P(V \geqslant 2) P(Z \leqslant 1)}{P(Y \geqslant 2)[1 - P(Z \geqslant 2 | Y \geqslant 2)]} = \frac{\alpha(m-x)[1-\alpha(x)]}{\alpha(m)[1-\alpha(x)/\alpha(m)]}$$

$$= \begin{cases} \dfrac{\alpha(m-x)[1-\alpha(x)]}{\alpha(m) - \alpha(x)} & \text{for } m-x \geqslant 2, \\ 0 & \text{for } m-x = 1. \end{cases} \tag{3.3}$$

Before describing the procedure, consider the following simple fact that was given in a more general setting by Sobel and Groll [3].

FACT 3.1 *Let a be equal to* 0 *and b be any integer in Definition 2.1 of a generalized binary group test. Let* $S1$ *and* $S2$ *be two disjoint binomial sets of objects. If a test* $T(0,b;S1 \cup S2)$ *produces a "bad" outcome and another test* $T(0,b;S1)$ *also produces a "bad" reading, then the conditional distribution of objects in* $S2$ *is the same as the original binomial distribution.*

If at the beginning of some step, the conflicted and mystery sets are both empty (i.e., we are faced only with a binomial set), we call this an "H-situation"; if the conflicted set is not empty and the mystery set is empty, we call this a "G-situation"; if the mystery set is not empty, we call this an "F-situation." By "resolving" a particular situation we mean that we classify the unclassified portion of the population when we are faced with that particular situation. Let $H(n)$ be the expected number of group tests needed to resolve an H-situation when the binomial set is of size n. Let $G(m,n)$ be the expected number of group tests needed to resolve a G-situation when the conflicted set is of size m and the binomial set is of size $n-m$. Let $x(H)$ and $x(G)$ be integers to be optimized later. The proposed group testing procedure is as follows:

CNC group testing procedure. If we have an H-situation, then a subset of the binomial set of size $x(H)$ is tested; if we have a G-situation, then a subset of the conflicted set of size $x(G)$ is tested; if we have an F-situation, then the entire mystery set is tested. After each test, the binomial, conflicted, and mystery sets are updated using the outcome of the test and Fact 3.1.

The operation of the protocol can be expressed as a pair of recursive equations along with a set of boundary conditions. These equations are:

$$H(n) = 1 + \min_{1 \leqslant x \leqslant n} \{[1 - \alpha(x)]H(n-x) + \alpha(x)G(x,n)\} \qquad \text{for } n \geqslant 2, \tag{3.4}$$

$$\begin{aligned} G(m,n) = 1 + \min_{1 \leqslant x \leqslant m-1} \{ & \gamma(x,m)G(x,n) \\ & + [1 - \gamma(x,m)][1 + \delta(x,m)G(m-x,n-x) \\ & + (1 - \delta(x,m))H(n-m)]\} \quad \text{for } n \geqslant m \geqslant 3. \end{aligned} \tag{3.5}$$

The boundary conditions are:

$$H(0) = 0, \tag{3.6}$$

$$G(0,n) = H(n), \tag{3.7}$$

$$G(2,2) = 2, \tag{3.8}$$

$$G(2,n) = 2+H(n-2) \qquad \text{for } n \geqslant 3. \tag{3.9}$$

Let integers $x(H(n))$ and $x(G(m,n))$ be the values of x that achieve the minimization in (3.4) and (3.5). These integers define the procedure implicitly. They can be found by solving (3.4) and (3.5) recursively, starting with the boundary conditions (3.6)-(3.9). If m is greater than 1, it is assumed that a subset of the conflicted set of size x with $1 \leqslant x \leqslant m-1$ will be tested without mixing it with items from the binomial set. This is referred to as the "nonmixing" rule. (Note: Consider a G-situation where the conflicted set is of size two. In such a situation, the defectiveness of the two members of the conflicted set can be deduced without any further tests. However, in some applications this observation is not sufficient to satisfy the objectives of that particular application. For example, in the communication problem, the objective is to transmit the active users' messages. In such application, knowing that the pair of users are both active is not sufficient, and the algorithm has to use two "steps" to transmit their messages. The above discussion explains why $G(2,2)=2$ and similarly explains the 2 in expression for $G(2,n)$. Conversely, there are other applications where the knowledge of defectiveness of items is sufficient; in that case $G(2,2)$ should be zero and the 2 in the expression of $G(2,n)$ should be eliminated.)

The algorithm is initiated in an H-situation where the binomial set is of size M. Hence, the performance of this algorithm is measured by $H(M)$ which is given by (3.4). Figure 1 shows $H(M)$ as a function of p for $M=64$.

The rest of this section is devoted to a brief discussion of some of the properties of the proposed algorithm. Let $p^*_{CNC}(M)$ be the value of p where the above algorithm adapts itself to testing groups of size one only. One of the interesting properties of this algorithm is that $p^*_{CNC}(M) = p^*_{CNC} = 1/\sqrt{2}$ independently of the original population size [2]. It has also been shown that this algorithm is the optimum strategy for $p \geqslant 1/\sqrt{2}$; i.e., if $p \geqslant 1/\sqrt{2}$, the best strategy in the sense of minimizing the expected number of group tests is to test groups of size one only. The proof is given in [2] in terms of the communication problem. Upper-bounds to the performance of the (yet unknown) optimum strategy for CNC problem is a communication setting is given in [2]. The above algorithm also outperforms algorithms based on binary tree search techniques originally designed for the communication problem. A comparison of such tree search algorithms and the above algorithm can be found in [6].

4. Success/Failure group testing. Consider the SF group test that was introduced in Section 2. The SF test was also motivated by a problem in the area of random-access communication systems [2]. A SF group testing problem is defined as follows:

DEFINITION 4.1. A Success/Failure (SF) group testing problem is concerned with partitioning a population of size M into subsets such that each subset contains at most one defective item. We associate independent Bernoulli $(p, q=1-p)$ random variables with each of the M objects, where p is the probability that an object is defective. We have at our disposal a Success/Failure generalized binary group test, $T(1,1;G)$. The goal is to design efficient algorithms for correctly partitioning the population.

Again, for finite M, the efficiency is defined in the sense of minimizing the expected number of group tests.

We now propose and analyze an algorithm for the SF group testing. Consider the setting where p is close to zero. Then, with high probability, a population of objects does not include any defective. However, since the SF test is not capable of distinguishing between a non-defective group and a conflicted group, a "bad" reading could be misleading. This inefficiency suggests the introduction of an $(M+1)st$ item that we shall refer to as the "auxiliary" item. The auxiliary item is a defective item that is always exclusively included (by the tester) in the first group to be tested, i.e., when the algorithm is initiated. Therefore, when none of the "real" items are defective, the auxiliary item is the only defective item. Hence, the auxiliary item transforms the SF group test into a conventional (SN) group test for the first step of the algorithm.

We shall refer to the first step of the algorithm, when all M items are in a binomial set, as an H-situation. If at some later step all of the unclassified items are in a binomial set, we call it an "$\tilde{H}$-situation". Let $H(M)$ be the expected number of group tests needed to resolve an H-situation and let $\tilde{H}$(n) be the expected number of group tests needed to resolve an $\tilde{H}$-situation when the binomial set is of size n. Let $x(H)$ and $x(\tilde{H})$ be integers to be optimized later. The proposed algorithm is as follows:

SF group testing procedure. If an H-situation exists, test a subset of the binomial set of size $x(H)$. If this results in a "bad" reading, then the same $x(H)$ items are tested individually in the next $x(H)$ tests, after which the size of the binomial set is reduced by $x(H)$. If, however, a "good" reading is obtained, the subset does not require any more tests and the size of the binomial set is again reduced by $x(H)$. If an $\tilde{H}$-situation exists, then proceed as above except use $x(\tilde{H})$ instead of $x(H)$.

The role of the auxiliary item becomes clear by studying the following pair of recursive equations describing the operation of the algorithm. These equations are:

$$H(M) = 1 + \min_{1 \leqslant x \leqslant M} \{q^x \tilde{H}(M-x) + (1-q^x)[x+\tilde{H}(M-x)]\}, \tag{4.1}$$

$$\tilde{H}(n) = 1 + \min_{1 \leqslant x \leqslant n} \{xpq^{x-1}\tilde{H}(n-x) + (1-xpq^{x-1})[x+\tilde{H}(n-x)]\}, \tag{4.2}$$

with the boundary conditions

$$H(0) = \tilde{H}(0) = 0. \tag{4.3}$$

Equation (4.1) represents the operation of the algorithm for the first step of the algorithm where the auxiliary item is present. The next equation is for the remaining steps where the auxiliary item is not being used.

The performance of this algorithm is shown in Figure 1 for M=64 as a function of p. Note that the protocol performs better than testing individual items for small values of p. Let $p^*_{SF}(M)$ be the value of p at which the algorithm tests individual items. The value of $H(M)$ for this algorithm is equal to M+1 for $p \geqslant p^*_{SF}(M)$, which is one more than is required for testing individual items. This is due to the one test used by the auxiliary user. Hence, a better strategy would be to use the proposed algorithm for $0 \leqslant p \leqslant \hat{p}_{SF}(M)$ and then test individual items without the help of the auxiliary item for $p \geqslant \hat{p}_{SF}(M)$, where $\hat{p}_{SF}(M)$ is the value of p at which $W(M)=M$. In contrast to the CNC and the conventional group testing, $p^*_{SF}(M)$ depends on the value of M. Note that a slightly better procedure can be obtained by deploying the auxiliary item throughout the entire procedure; however, this may bring some hardship for the tester. Upper-bounds to the performance of the (yet unknown) optimum strategy for SN problem in a communication setting is given in [2].

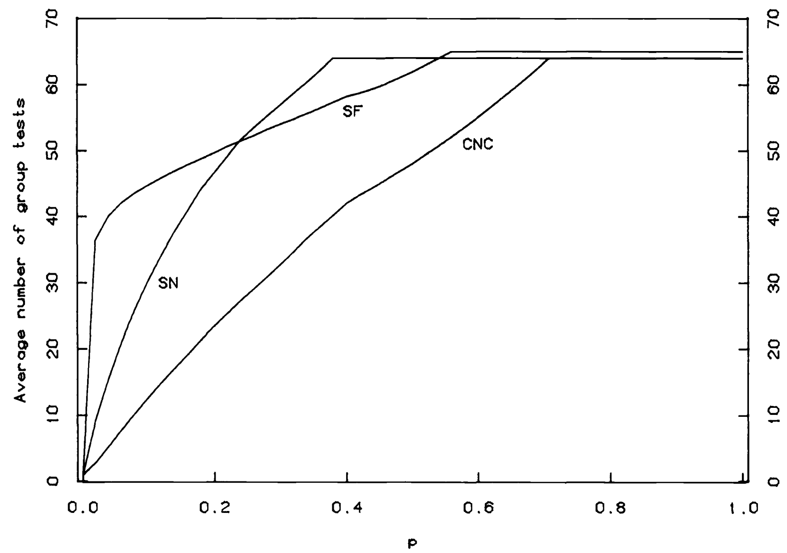


FIG. 1. *Performance of the group testing procedures.*

REFERENCES

[1] ROBERT DORFMAN, *Detection of defective members of large populations*, Ann. of Math. Stat., 28 (1943), pp. 436-440.

[2] NADER MEHRAVARI, *Random multiple-access communication with binary feedback*, PhD Thesis, School of Elect. Engr., Cornell Univ., Ithaca, New York, August 1982.

[3] MILTON SOBEL AND PHYLLIS A. GROLL, *Group testing to eliminate efficiently all defectives in a binomial sample*, Bell System Technical Journal, 38 (1959), pp. 1179-1253.

[4] ANDREW STERRETT, *On the detection of defective members of large populations*, Ann. of Math. Stat., 28 (1957), pp. 1033-1036.

[5] PETER UNGAR, *The cutoff point for group testing*, Comm. Pure Appl. Math., 13 (1960), pp. 49-54.

[6] T. BERGER, N. MEHRAVARI, D. TOWSLEY, AND J. WOLF, *Random multiple-access communication and group testing*, IEEE Trans. Comm., COMM-23 (1984), pp. 769-779.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

001

ON THE EIGENVALUE PROBLEM FOR A CLASS OF BAND MATRICES INCLUDING THOSE WITH TOEPLITZ INVERSES*

WILLIAM F. TRENCH†

Abstract. We study the eigenvalue problem for a class $\mathcal{H}$ of band matrices which includes as a proper subclass all band matrices with Toeplitz inverses. Toeplitz matrices of this kind occur, for example, as autocorrelation matrices of purely autoregressive stationary time series. A formula is given for the characteristic polynomial $p_n(\lambda)$ of an nth order matrix H_n in $\mathcal{H}$, with bandwidth $k+1 \leqq n$, as the ratio of $k \times k$ determinants whose entries are polynomials in the zeros of a certain kth degree polynomial which is independent of n and has one coefficient which depends upon λ. The formula permits the evaluation of $p_n(\lambda)$ by means of a computation with complexity independent of n. Also given is a formula for the eigenvectors in terms of these zeros and k coefficients which can be obtained by solving a $k \times k$ homogeneous system.

AMS(MOS) subject classification. 15A18

1. Introduction. We consider the eigenvalue problem for the class $\mathcal{H}$ of matrices

$$H_n = (h_{ijn})_{i,j=0}^{n-1}, \tag{1}$$

defined as follows. Let

$$A(z) = \sum_{\nu=0}^{r} a_\nu z^\nu, \qquad B(z) = \sum_{\mu=0}^{s} b_\mu z^\mu,$$

where

$$a_0 b_0 \neq 0 \quad \text{and} \quad r+s=k<n, \tag{2}$$

and $\{h_{ijn}\}$ are defined by the generating functions

$$H_{in}(z) = \sum_{j=0}^{n-1} h_{ijn} z^j = \begin{cases} z^i A(z) \sum_{\mu=0}^{i} b_\mu z^{-\mu}, & 0 \leqq i \leqq s-1, \\ z^i A(z) B(1/z), & s \leqq i \leqq n-r-1, \\ z^i B(1/z) \sum_{\nu=0}^{n-i-1} a_\nu z^\nu, & n-r \leqq i \leqq n-1. \end{cases}$$

Explicitly,

$$h_{ijn} = c_{j-i} - \sum_{\nu=i+1}^{s} a_{j-i+\nu} b_\nu - \sum_{\mu=n-i}^{r} b_{i-j+\mu} a_\mu, \qquad 0 \leqq i, j \leqq n-1, \tag{3}$$

if we define

$$a_l = 0 \text{ if } l > r \text{ or } l < 0, \quad b_l = 0 \text{ if } l > s \text{ or } l < 0, \quad \sum_q^p = 0 \text{ if } q > p, \tag{4}$$

$$c_\nu = 0 \quad \text{if } \nu > r \text{ or } \nu < -s, \tag{5}$$

and

$$C(z) = A(z)B(1/z) = \sum_{\nu=-s}^{r} c_\nu z^\nu. \tag{6}$$

* Received by the editors March 1, 1984, and in revised form December 5, 1984.

† Department of Mathematics and Computer Science, Drexel University, Philadelphia, Pennsylvania 19104.

The class $\mathcal{H}$ is connected with Toeplitz matrices; i.e., matrices of the form

$$T_n = (\phi_{j-i})_{i,j=0}^{n-1}.$$

From (3) and (4),

$$h_{ijn} = c_{j-i} \quad \text{if} \begin{cases} 0 \leqq i \leqq s-1 \text{ and } r \leqq j \leqq n-1, \\ \text{or } s \leqq i \leqq n-r-1, \\ \text{or } n-r \leqq i \leqq n-1 \text{ and } 0 \leqq j \leqq n-s-1; \end{cases} \tag{7}$$

thus, H_n is *quasi-Toeplitz* (a term used in [7]) in that h_{ijn} is a function of $j-i$ alone except in the $s \times r$ submatrix in the upper left corner of H_n and the $r \times s$ submatrix in the lower right corner. Moreover, H_n is *banded*; i.e.,

$$h_{ijn} = 0 \quad \text{if } j-i > r \text{ or } i-j > s, \tag{8}$$

from (3), (4), and (5).

Matrices in the class $\mathcal{H}$ have been encountered by the author [10] in connection with prediction of stationary time series, and by Greville [4], [5], [6], in connection with a smoothing problem. Greville and the author studied them in [7], and obtained results which can be summarized as follows.

THEOREM 1 (Greville-Trench). *The matrices H_n $(n > k)$ are invertible if and only if $A(z)$ and $z^sB(1/z)$ are relatively prime, in which case their inverses are the Toeplitz matrices*

$$H_n^{-1} = T_n = (\phi_{j-i})_{i,j=0}^{n-1}, \qquad n > k,$$

where $\{\phi_j\}$ is determined as follows: Obtain $[\phi_{s-1}, \phi_{s-2}, \cdots, \phi_{-r}]$ as the (unique) solution of the $k \times k$ system

$$\begin{aligned} &\text{(a)} \quad \sum_{\nu=0}^{r} a_\nu \phi_{j-\nu} = b_0^{-1}\delta_{j0}, \qquad 0 \leqq j \leqq s-1, \\ &\text{(b)} \quad \sum_{\mu=0}^{s} b_\mu \phi_{-j+\mu} = 0, \qquad 1 \leqq j \leqq r, \end{aligned} \tag{9}$$

and then compute

$$\phi_j = -a_0^{-1} \sum_{\nu=1}^{r} a_\nu \phi_{j-\nu}, \qquad j \geqq s, \tag{10}$$

and

$$\phi_{-j} = -b_0^{-1} \sum_{\mu=1}^{s} b_\mu \phi_{-j+\mu}, \qquad j > r. \tag{11}$$

Moreover, if H_n $(n > k)$ is a matrix of the form (1) *such that* (8) *holds and H_n^{-1} is a Toeplitz matrix, then $H_n \in \mathcal{H}$.*

Greville continued the investigation of these matrices in [2] and [3].

The main result of this paper reduces the evaluation of the characteristic polynomial $p_n(\lambda)$ of H_n to finding the zeros of the polynomial

$$P(z; \lambda) = \sum_{\mu=-s}^{r} c_\mu z^{\mu+s} - \lambda z^s \tag{12}$$

(which are obviously independent of n) and evaluating a kth order determinant whose entries are polynomials in these zeros. The complexity of this representation of $p_n(\lambda)$

depends only on k (cf. (2)), and is independent of n. Moreover, we give an explicit formula for the eigenvectors of H_n corresponding to a given eigenvalue, which depends on k coefficients that can be obtained by solving a kth order homogeneous system with complexity independent of n. The results are analogous to those obtained in [11] for Toeplitz band matrices

$$T_n = (c_{j-i})_{i,j=0}^{n-1}, \tag{13}$$

where $\{c_\nu\}$ satisfies (5) and $r+s=k<n$. However, the arguments needed here are considerably more complicated than those in [11].

Our results here are not restricted to the case where $A(z)$ and $z^s B(1/z)$ are relatively prime, so that H_n is invertible; however, this case is especially important, since Theorem 1 implies that the eigenvalue problems for invertible matrices in $\mathscr{H}$ and for Toeplitz matrices with banded inverses are equivalent. Although there is a large body of literature on inverting Toeplitz matrices and solving systems with Toeplitz matrices, little has been published on approaches to the eigenvalue problem for these matrices which take full advantage of their simple structure. (For examples, see Grunbaum [8], [9]; Dini and Capovani [1]; and Trench [11].)

2. Preliminary definitions and lemmas. We take the underlying field to be the complex numbers.

It can be seen from (7) and (8) that if r or s is zero, then H_n is a triangular Toeplitz matrix. Since the eigenvalue problem for such matrices is trivial, we assume henceforth that (2) holds, and also that

$$rsa_rb_s \neq 0. \tag{14}$$

Then $rsc_rc_{-s}\neq 0$, so $P(0;\lambda)\neq 0$.

It was shown in [11] that there are at most k values of λ for which $P(z;\lambda)$ has fewer than k distinct zeros. We call such points *critical points* of $P(z;\lambda)$. All other values of λ are *ordinary points.* For completeness, we phrase all definitions so as to include the case where λ is a critical point; however, for notational convenience we illustrate the definitions only for ordinary points.

DEFINITION 1. For a fixed λ, let $z_1,\cdots,z_q$ be the distinct zeros of (12) with multiplicities $m_1,\cdots,m_q$; thus,

$$q\leqq k,\quad m_j\geqq 1\ (1\leqq j\leqq k),\quad m_1+\cdots m_q=k.$$

If $Q_1(z),\cdots,Q_k(z)$ are given polynomials, define the k-vector function

$$w(z)=\operatorname{col}[Q_1(z),Q_2(z),\cdots,Q_k(z)],$$

and let Ω be the $k\times k$ matrix defined as follows: its first m_1 columns are $w^{(l)}(z_1)$ $(0\leqq l\leqq m_1-1)$; its next m_2 columns are $w^{(l)}(z_2)$ $(0\leqq l\leqq m_2-1)$; and so forth. Let V be the matrix resulting from this construction in the special case where $Q_i(z)=z^{i-1}$, $1\leqq i\leqq k$, and define

$$\Delta(\lambda)=\frac{\det\Omega}{\det V}. \tag{15}$$

Thus, if λ is an ordinary point of $P(z;\lambda)$, then $q=k$, $m_1=\cdots=m_k=1$,

$$\Omega=(Q_i(z_j))_{i,j=1}^{k}$$

and V is the Vandermonde matrix

$$V=(z_j^{i-1})_{i,j=1}^{k}.$$

It can be shown in general that

$$\det V = K_{ij} \prod_{i \leqq i < j \leqq q} (z_j - z_i)^{r_{ij}} \qquad (K_{ij} = \text{constant} \neq 0),$$

where the r_{ij}'s are positive integers (all ones if $q = k$); hence, $\det V \neq 0$.

We refrain from using the functional notation $z_i(\lambda)$ for the root z_i, since this would necessitate an irrelevant appeal to the theory of multiple-valued algebraic functions. Note that there is an ambiguity in the definitions of Ω and V, since the numbering of the roots is arbitrary; however, since renumbering $z_1, \cdots, z_q$ would simply permute the columns of both matrices in the same way, the ratio of the determinants in (15) is uniquely defined for each λ.

To avoid cumbersome two-dimensional displays in proofs which follow, we also denote the function $\Delta(\lambda)$ defined by (15) in the form

$$\Delta(\lambda) = |Q_1(z), \cdots, Q_k(z)|(\lambda); \tag{16}$$

thus, if λ is an ordinary point of $P(z; \lambda)$, then

$$|Q_1(z), \cdots, Q_k(z)|(\lambda) = \frac{\det (Q_i(z_j))_{i,j=1}^k}{\det (z_j^{i-1})_{i,j=1}^k}. \tag{17}$$

There is an abuse of notation here, since (17) is not a function of z as the symbol on the left appears to indicate; however, the convenience of the notation outweighs this drawback.

In the following we adopt the convention that the polynomial $p = 0$ has degree $-\infty$.

LEMMA 1. *Let $n_1, \cdots, n_k$ be nonnegative integers, and $m = \max \{n_1, \cdots, n_k\}$. Then the function*

$$|z^{n_1}, \cdots, z^{n_k}|(\lambda) \tag{18}$$

is a polynomial of degree $\leqq m - k + 1$.

This lemma was proved in [11], where the function in (18) was denoted by $q(\lambda; n_1, \cdots, n_k)$. The main result in [11] is that the characteristic polynomial of the Toeplitz band matrix T_n in (13) is

$$\det [\lambda I_n - T_n] = (-1)^{(r-1)n} c_r^n |1, z, \cdots, z^{s-1}, z^{n+s}, \cdots, z^{n+k-1}|(\lambda).$$

We will obtain an analogous result here for H_n.

LEMMA 2. *Let $Q_1(z), \cdots, Q_k(z)$ and $\Delta(\lambda)$ be as in Definition 1, and let $m = \max_{1 \leqq i \leqq k} \{\deg Q_i(z)\}$. Then $\Delta(\lambda)$ is a polynomial of degree $\leqq m - k + 1$. Moreover, $\Delta(\lambda) = 0$ for a given complex number λ if and only if there are constants $c_1, \cdots, c_k$ (not all zero) such that the polynomial*

$$Q(z) = c_1 Q_1(z) + \cdots + c_k Q_k(z) \tag{19}$$

is divisible by $P(z; \lambda)$. In particular, if $m \leqq k - 1$, so that $\Delta(\lambda) = C$ (constant), then $C = 0$ if and only if $Q_1(z), \cdots, Q_k(z)$ are linearly dependent.

Proof. If

$$Q_i(z) = \sum_{j=0}^{m} a_{ij} z^j, \qquad 1 \leqq i \leqq k,$$

then

$$\Delta(\lambda) = \sum a_{1j_1} \cdots a_{kj_k} |z^{j_1}, \cdots, z^{j_k}|(\lambda),$$

where the sum is over all $j_1, \cdots, j_k$ such that $0 \leqq j_1, \cdots, j_k \leqq m$, so $\Delta(\lambda)$ is a polynomial of degree $\leqq m-k+1$, by Lemma 1. From (15), we see that $\Delta(\lambda)=0$ if and only if the system

$$\Omega' Y=0 \quad ('=\text{transpose}) \tag{20}$$

has a nontrivial solution $Y=[c_1, \cdots, c_k]$. From the definition of Ω, it can be seen that (20) is equivalent to

$$Q^{(l)}(z_j)=0, \qquad 0\leqq l\leqq m_j-1, \quad 1\leqq j\leqq q, \tag{21}$$

with $Q(z)$ as in (19). But (21) holds if and only if $P(z;\lambda)$ divides $Q(z)$. This implies the stated conclusions.

LEMMA 3. *With the assumptions of Lemma* 2, *suppose that*

$$Q_i(z)=\alpha z^m+\cdots$$

for some $m\geqq k$ *and* i *in* $\{1, \cdots. k\}$, *and that*

$$\deg Q_j(z)<m \quad \text{if } j\neq i. \tag{22}$$

Then $\Delta(\lambda)$ *as defined in* (16) *can be written as*

$$\Delta(\lambda)=\alpha\lambda c_r^{-1}|Q_1(z), \cdots, Q_{i-1}(z), z^{m-r}, Q_{i+1}(z), \cdots, Q_k(z)|(\lambda)+O(\lambda^{m-k}), \tag{23}$$

where $O(\lambda^{m-k})$ *denotes a polynomial of degree* $\leqq m-k$.

Proof. From elementary properties of determinants,

$$\Delta(\lambda)=|\cdots, \alpha z^m, \cdots|(\lambda)+|\cdots, Q_i(z)-\alpha z^m, \cdots|(\lambda), \tag{24}$$

where the first "$\cdots$" denotes "$Q_i(z), \cdots, Q_{i-1}(z)$" and the second "$\cdots$" denotes "$Q_{i+1}(z), \cdots, Q_k(z)$" throughout this proof. From (22) and Lemma 2, the second term on the right of (24) is $O(\lambda^{m-k})$; hence,

$$\Delta(\lambda)=\alpha|\cdots, z^m, \cdots|(\lambda)+O(\lambda^{m-k}). \tag{25}$$

Now we use the identity

$$z^m=c_r^{-1}\left[\lambda z^{m-r}-\sum_{\mu=-s}^{r-1} c_\mu z^{\mu+m-r}+z^{m-k}P(z,\lambda)\right]$$

(see (12)) to write

$$\begin{aligned} |\cdots, z^m, \cdots|(\lambda) &= \lambda c_r^{-1}|\cdots, z^{m-r}, \cdots|(\lambda) \\ &\quad -\sum_{\mu=-s}^{r-1} c_\mu c_r^{-1}|\cdots, z^{\mu+m-r}, \cdots|(\lambda) \\ &\quad +c_r^{-1}|\cdots, z^{m-k}P(z,\lambda), \cdots|(\lambda). \end{aligned} \tag{26}$$

From Definition 1, the ith row of the determinant in the numerator of

$$|\cdots, z^{m-k}P(z,\lambda), \cdots|(\lambda)$$

consists entirely of zeros, so the last term on the right of (26) vanishes. Lemma 2 and (22) imply that the sum on the right of (26) is $O(\lambda^{m-k})$. Therefore,

$$|\cdots, z^m, \cdots|(\lambda)=\lambda c_r^{-1}|\cdots, z^{m-r}, \cdots|(\lambda)+O(\lambda^{m-k}).$$

This and (25) imply (23).

We now establish a useful connection between the eigenvalue problem for H_n and a boundary value problem for a related difference equation. Let

$$U=\operatorname{col}[u_0,\cdots,u_{n-1}]$$

and

$$V=H_nU=\operatorname{col}[v_0,\cdots,v_{n-1}]. \tag{27}$$

Then,

$$v_i=\sum_{j=0}^{n-1}h_{ijn}u_j=\sum_{j=-i}^{n-i-1}h_{i,j+i,n}u_{j+i},\qquad 0\leqq i\leqq n-1.$$

Therefore, from (3),

$$v_i=\sum_{j=-i}^{n-i-1}\left[c_j-\sum_{\nu=i+1}^{s}a_{j+\nu}b_\nu-\sum_{\mu=n-i}^{r}b_{-j+\mu}a_\mu\right]u_{j+i},\qquad 0\leqq i\leqq n-1. \tag{28}$$

If $s\leqq i\leqq n-r-1$, then the sums with respect to μ and ν both vanish, and (28) reduces to

$$v_i=\sum_{j=-s}^{r}c_ju_{j+i}. \tag{29}$$

For our purposes it is convenient to have this equation hold for $0\leqq i\leqq n-1$, but this is impossible as things stand, since (29) would then involve the undefined quantities $u_{-s},\cdots,u_{-1}$ and $u_n,\cdots,u_{n+r-1}$. This defect can be remedied by defining *extrapolated components* for the vector U, as in the following lemma.

LEMMA 4. *The components of V in* (27) *are given by* (29) *for $0\leqq i\leqq n-1$ if and only if the extrapolated components $u_{-s},\cdots,u_{-1}$ and $u_n,\cdots,u_{n+r-1}$ satisfy the equations*

$$\sum_{l=0}^{r}a_lu_{l-p}=0,\qquad 1\leqq p\leqq s, \tag{30}$$

and

$$\sum_{l=0}^{s}b_lu_{n+p-l-1}=0,\qquad 1\leqq p\leqq r. \tag{31}$$

Proof. The extrapolated components are uniquely defined by $u_0,\cdots,u_{n-1}$ and (30) and (31). They can be computed recursively from the equations

$$u_{-p}=-a_0^{-1}\sum_{l=1}^{r}a_lu_{l-p},\qquad 1\leqq p\leqq s, \tag{32}$$

and

$$u_{n+p-1}=-b_0^{-1}\sum_{l=1}^{s}b_lu_{n+p-l-1},\qquad 1\leqq p\leqq r. \tag{33}$$

We have already verified (29) for $s\leqq i\leqq n-r-1$. If $0\leqq i\leqq s-1$, then (28) reduces to

$$v_i=\sum_{j=-i}^{r}\left(c_j-\sum_{\nu=i+1}^{s}a_{j+\nu}b_\nu\right)u_{j+i} \tag{34}$$

because of (2), (4), and (5). From (6),

$$c_j=\sum_{\nu=i+1}^{s}a_{j+\nu}b_\nu,\qquad -s\leqq j\leqq -i-1,$$

(since $a_{j+\nu}=0$ if $j+\nu<0$); hence, we can change the lower limit of summation in (34) to $j=-s$ for any choice of $u_{-s},\cdots,u_{-1}$. Therefore,

$$v_i=\sum_{j=-s}^{r} c_j u_{j+i}-\sum_{j=-s}^{r}\left(\sum_{\nu=i+1}^{s} a_{j+\nu}b_\nu\right)u_{j+i}, \qquad 0\leqq i\leqq s-1. \tag{35}$$

The double sum in (35) can be rewritten as

$$\begin{aligned}\sum_{\nu=i+1}^{s} b_\nu \sum_{j=-s}^{r} a_{j+\nu}u_{j+i} &= \sum_{\nu=i+1}^{s} b_\nu \sum_{l=\nu-s}^{\nu+r} a_l u_{l+i-\nu}\\ &= \sum_{\nu=i+1}^{s} b_\nu \sum_{l=0}^{r} a_l u_{l-(\nu-i)}, \qquad 0\leqq i\leqq s-1\end{aligned}$$

(since $a_l=0$ if $l<0$ or $l>r$). From this it can be seen that (29) holds for $0\leqq i\leqq s-1$ if and only if (30) holds.

Now suppose $n-r\leqq i\leqq n-1$. Then (28) reduces to

$$v_i=\sum_{j=-s}^{n-i-1}\left(c_j-\sum_{\mu=n-i}^{r} b_{-j+\mu}a_\mu\right)u_{j+i}. \tag{36}$$

Since (6) implies that

$$c_j=\sum_{\mu=n-i}^{r} b_{-j+\mu}a_\mu, \qquad n-i\leqq j\leqq r$$

(recall that $b_{-j+\mu}=0$ if $-j+\mu<0$), we can change the upper limit of summation in (36) to $j=r$ for any choice of $u_n,\cdots,u_{n+r-1}$. Therefore,

$$v_i=\sum_{j=-s}^{r} c_j u_{j+i}-\sum_{j=-s}^{r}\left(\sum_{\mu=n-i}^{r} b_{-j+\mu}a_\mu\right)u_{j+i}, \qquad n-r\leqq i\leqq n-1.$$

The double sum on the right can be rewritten as

$$\begin{aligned}\sum_{\mu=n-i}^{r} a_\mu \sum_{j=-s}^{r} b_{-j+\mu}u_{j+i} &= \sum_{\mu=n-i}^{r} a_\mu \sum_{l=\mu-r}^{\mu+s} b_l u_{\mu-l+i}\\ &= \sum_{\mu=n-i}^{r} a_\mu \sum_{l=0}^{s} b_l \mu_{\mu-l+i}, \qquad n-r\leqq i\leqq n-1\end{aligned}$$

(since $b_l=0$ if $l<0$ or $l>s$). From this it can be seen that (29) holds for $n-r\leqq i\leqq n-1$ if and only if (31) holds.

Lemma 4 obviously implies the following lemma.

LEMMA 5. *A complex number λ is an eigenvalue of H_n if and only if there are complex numbers*

$$u_{-s},\cdots,u_{n+r-1}, \tag{37}$$

not all zero, which satisfy the difference equation

$$\sum_{j=-s}^{r} c_j u_{j+i}=\lambda u_i, \qquad 0\leqq i\leqq n-1, \tag{38}$$

and the boundary conditions (30) *and* (31). *In this case the vector*

$$U=\operatorname{col}[u_0,\cdots,u_{n-1}] \tag{39}$$

is an eigenvector of H_n corresponding to λ.

It is important to observe that if the sequence (37) satisfies these hypotheses, then U in (39) is nonzero, since if $u_0 = \cdots = u_{n-1} = 0$, then (32) and (33) imply that the remaining elements in (37) vanish.

3. The main results. In the following $(x)^{(l)}$ is the factorial polynomial:

$$(x)^{(0)} = 1, \qquad (x)^{(l)} = x(x-1)\cdots(x-l+1), \quad l \geqq 1.$$

THEOREM 2. *Let λ satisfy the assumptions of Definition 1, and let Ω_n be the $k\times k$ matrix which results from the construction specified in Definition 1 when*

$$Q_i(z) = \begin{cases} z^{i-1}A(z), & 1 \leqq i \leqq s, \\ z^{n+i-1}B(1/z), & s+1 \leqq i \leqq k; \end{cases}$$

thus

$$\Omega_n = \begin{bmatrix} A(z_1) & A(z_2) & \cdots & A(z_k) \\ \vdots & \vdots & & \vdots \\ z_1^{s-1}A(z_1) & z_2^{s-1}A(z_2) & \cdots & z_k^{s-1}A(z_k) \\ z_1^{n+s}B(1/z_1) & z_2^{n+s}B(1/z_2) & \cdots & z_k^{n+s}B(1/z_k) \\ \vdots & \vdots & & \vdots \\ z_1^{n+k-1}B(1/z_1) & z_2^{n+k-1}B(1/z_2) & \cdots & z_k^{n+k-1}B(1/z_k) \end{bmatrix} \tag{40}$$

if λ is an ordinary point of $P(z;\lambda)$. Then λ is an eigenvalue of H_n if and only if Ω_n is singular, in which case the components of the eigenvector (39) *are given by*

$$u_i = \sum_{j=1}^{q} \sum_{\nu=0}^{m_j-1} \alpha_{\nu j}(s+i)^{(\nu)} z_j^{s+i-\nu} \tag{41}$$

for $0 \leqq i \leqq n-1$, where the vector

$$X = \operatorname{col}[\alpha_{01}, \cdots, \alpha_{m_1-1,1}, \alpha_{02}, \cdots, \alpha_{m_2-1,2}, \cdots, \alpha_{0q}, \cdots, \alpha_{m_q-1,q}] \tag{42}$$

is a nontrivial solution of the $k\times k$ system

$$\Omega_n X = 0. \tag{43}$$

(*Note that* (41) *and* (42) *can be written more simply as*

$$u_i = \sum_{j=1}^{k} \alpha_j z_j^{s+i}, \qquad 0 \leqq i \leqq n-1,$$

and

$$X = \operatorname{col}[\alpha_1, \alpha_2, \cdots, \alpha_k]$$

if λ is an ordinary point of $P(z;\lambda)$.)

Proof. We use Lemma 5. The general solution of the difference equation (38) is of the form (41) for $-s \leqq i \leqq n+r-1$. (See the proof of Theorem 1 in [11].) Substituting (41) into (30) and summing first on l yields

$$\sum_{j=1}^{q} \sum_{\nu=0}^{m_j-1} \alpha_{\nu j} \sum_{l=0}^{r} a_l(s+l-p)^{(\nu)} z_j^{s+l-p-\nu} = 0, \qquad 1 \leqq p \leqq s.$$

This is equivalent to

$$\sum_{j=1}^{q} \sum_{\nu=0}^{m_j-1} \alpha_{\nu j}[(z^{s-p}A(z))^{(\nu)}|_{z=z_j}] = 0, \qquad 1 \leqq p \leqq s. \tag{44}$$

By a similar argument, substituting (41) into (31) yields

$$(45)\qquad \sum_{j=1}^{q}\sum_{\nu=0}^{m_j-1} \alpha_{\nu j}[(z^{s+n+p-1}B(1/z))^{(\nu)}|_{z=z_j}]=0, \qquad 1\leqq p\leqq r.$$

Since (44) and (45) together are equivalent to (43), the conclusion follows.

If we let $\Omega=\Omega_n$ in (15), then $\Delta(\lambda)$ in (16) becomes

$$(46)\qquad \Delta_n(\lambda)=|A(z),\cdots,z^{s-1}A(z),z^{n+s}B(1/z),\cdots,z^{n+k-1}B(1/z)|(\lambda),$$

which is a polynomial of degree $\leqq n$, by Lemma 2. Since Ω_n is singular if and only if $\Delta_n(\lambda)=0$, Theorem 2 clearly implies that there is a connection between $\Delta_n(\lambda)$ and the characteristic polynomial

$$(47)\qquad p_n(\lambda)=\det[\lambda I_n-H_n].$$

The following results make this connection precise.

THEOREM 3. *If $A(z)$ and $z^sB(1/z)$ are relatively prime, then*

$$(48)\qquad p_n(\lambda)=(-1)^{(r-1)n}R^{-1}c_r^n\Delta_n(\lambda), \qquad n>k,$$

where R is the (nonzero) value of the $k\times k$ determinant with rows $i=1,\cdots,k$ as follows:

(a) *For $1\leqq i\leqq s$, there are $i-1$ zeros, then $a_0,\cdots,a_r$, then $s-i$ zeros.*

(b) *For $s+1\leqq i\leqq k$, there are $i-s-1$ zeros, then $b_s,\cdots,b_0$, then $k-i$ zeros.*

Proof. Although $p_n(\lambda)$ has meaning only if $n>k$, $\Delta_n(\lambda)$ is defined for all $n\geqq 0$. We first prove by induction that

$$(49)\qquad \Delta_n(\lambda)=(-1)^{(r-1)n}Rc_r^{-n}\lambda^n+g_n(\lambda), \qquad n\geqq 0,$$

where $\deg g_n(\lambda)<n$. It suffices to consider only the case where λ is an ordinary point of $P(z;\lambda)$, since there are at most k critical values of λ, and we already know that $\Delta_n(\lambda)$ is a polynomial of degree $\leqq n$.

From (40), $\Omega_0=WV$, where V is the Vandermonde matrix of Definition 1 and $\det W=R$. This implies (49) for $n=0$, with $g_0=0$. To see that $R\neq 0$, suppose $R=0$. Then (40) with $n=0$ and the last sentence of Lemma 2 imply that

$$A(z),\cdots,z^{s-1}A(z),z^sB(1/z),\cdots,z^{k-1}B(1/z)$$

are linearly dependent. Therefore, there are polynomials $f(z)$ and $g(z)$, not identically zero, such that $\deg f(z)<s$, $\deg g(z)<r$, and $f(z)A(z)=g(z)z^sB(1/z)$. By an argument in [12, § 27], this implies that $A(z)$ and $z^sB(1/z)$ have a nonconstant common factor, which contradicts our assumption.

We now complete the proof of (49) by showing that

$$(50)\qquad \Delta_{n+1}(\lambda)=(-1)^{r-1}c_r^{-1}\lambda\Delta_n(\lambda)+O(\lambda^n), \qquad n\geqq 0,$$

where $O(\lambda^n)$ denotes a polynomial of degree $\leqq n$. From (46) with n replaced by $n+1$,

$$\Delta_{n+1}(\lambda)=|\cdots,z^{n+s+1}B(1/z),\cdots,z^{n+k}B(1/z)|(\lambda),$$

where the first "$\cdots$" denotes "$A(z),\cdots,z^{s-1}A(z)$" throughout this proof. The polynomial of highest degree appearing in the definition of $\Delta_{n+1}(\lambda)$ is $z^{n+k}B(1/z)$; hence, Lemma 3 implies that

$$(51)\quad \Delta_{n+1}(\lambda)=b_0\lambda c_r^{-1}|\cdots,z^{n+s+1}B(1/z),\cdots,z^{n+k-1}B(1/z),z^{n+s}|(\lambda)+O(\lambda^n),$$

where $z^{n+s+1}B(1/z), \cdots, z^{n+k-1}B(1/z)$ is absent if $r=1$. We rewrite (51) as

$$\begin{aligned}\Delta_{n+1}(\lambda) &= (-1)^{r-1}\lambda c_r^{-1}|\cdots, b_0 z^{n+s}, z^{n+s+1}B(1/z), \cdots, z^{n+k-1}B(1/z)|(\lambda)+O(\lambda^n)\\ &= (-1)^{r-1}\lambda c_r^{-1}[\Delta_n(\lambda)+\Gamma_n(\lambda)]+O(\lambda^n), \qquad (52)\end{aligned}$$

where

$$\Gamma_n(\lambda) = |\cdots, z^{n+s}(b_0 - B(1/z)), z^{n+s+1}B(1/z), \cdots, z^{n+k-1}B(1/z)|(\lambda). \tag{53}$$

(See (46).)

We will now show that

$$\Gamma_n(\lambda) = O(\lambda^{n-1}). \tag{54}$$

If $r=1$, then $k=s+1$ and (53) reduces to

$$\Gamma_n(\lambda) = |\cdots, z^{n+s}(b_0 - B(1/z))|(\lambda):$$

so Lemma 2 implies (54). If $r>1$, then successively applying Lemma 3 to (53) $r-1$ times yields

$$\begin{aligned}\Gamma_n(\lambda) &= (b_0\lambda c_r^{-1})^{r-1}|\cdots, z^{n+s}(b_0 - B(1/z)), z^{n+s-r+1}, \cdots, z^{n+s-1}|(\lambda)\\ &= (b_0\lambda c_r^{-1})^{r-1}\sum_{\mu=1}^{s} b_\mu|\cdots, z^{n+s-\mu}, z^{n+s-r+1}, \cdots, z^{n+s-1}|(\lambda).\end{aligned}$$

The terms in this sum are identically zero for $1 \leqq \mu \leqq \max(s, r-1)$ (since they are essentially determinants with two identical rows), and $O(\lambda^{n-r})$ for $r \leqq \mu \leqq s$ (by Lemma 2). This implies (54). Since (52) and (54) imply (50), this completes the proof of (49).

Now (49) implies that the polynomial

$$\tilde{p}_n(\lambda) = (-1)^{(r-1)n}R^{-1}c_r^n\Delta_n(\lambda)$$

is monic and of exact degree n, as is the characteristic polynomial $p_n(\lambda)$ in (47). From Theorem 2, $\tilde{p}_n(\lambda)$ and $p_n(\lambda)$ have the same zeros; therefore certainly $\tilde{p}_n(\lambda)=p_n(\lambda)$ if H_n has n distinct eigenvalues. There remains the possibility that H_n has only m $(<n)$ distinct eigenvalues and

$$\tilde{p}_n(\lambda) = (\lambda-\lambda_1)^{r_1}\cdots(\lambda-\lambda_m)^{r_m}, \quad p_n(\lambda) = (\lambda-\lambda_1)^{s_1}\cdots(\lambda-\lambda_m)^{s_m}$$

with $r_i \neq s_i$ for some i; however, this possibility can be excluded by a continuity argument of the kind given in [11].

Theorems 1 and 3 yield the following result, which makes explicit the connection between our results and the eigenvalue problem for Toeplitz matrices with band inverses.

THEOREM 4. *Suppose $A(z)$ and $B(z)$ satisfy* (2) *and* (14), *and $A(z)$ and $z^sB(1/z)$ are relatively prime. Let T_n and $\{\phi_r\}$ be as in Theorem* 1. *Then the characteristic polynomial of T_n is given by*

$$\det[\lambda I_n - T_n] = [\Delta_n(0)]^{-1}\lambda^n\Delta_n(1/\lambda).$$

Moreover, if λ is an eigenvalue of T_n, then the corresponding eigenvectors (39) *can be obtained as in Theorem* 2.

Our results have specific applications to statistics in the case where

$$B(z) = A^*(z) = \sum_{\nu=0}^{r} \bar{a}_\nu z^\nu,$$

so that the matrices H_n $(n>2r)$ are Hermitian. Greville [3] has shown that in this case

H_n is positive definite for all $n > 2r$ if and only if the zeros of $A(z)$ are all outside the unit circle, or positive semidefinite if and only if none are inside the unit circle. He also obtained results on the spectral radii of the matrices $\{H_n\}$.

If the roots of $A(z)$ are all outside the unit circle, then $A(z)$ and $z^r A^*(1/z)$ are relatively prime. It can be shown in this case that the sequence $\{\phi_r\}$ defined by (9), (10), and (11) (with $b_\nu = \bar{a}_\nu$) is proportional to the autocorrelation sequence of the purely autoregressive weakly stationary time series $\{y_m\}$ defined by the stochastic difference equation

$$a_0 y_m + a_1 y_{m-1} + \cdots + a_r y_{m-r} = x_m, \qquad -\infty < m < \infty,$$

where $\{x_m\}$ is uncorrelated and weakly stationary.

The formula (48) is clearly invalid if $A(z)$ and $z^s B(1/z)$ have a nonconstant common factor, since R is the resultant of $z^r A(1/z)$ and $B(z)$, which would also have a nonconstant common factor, and therefore $R = 0$ [12, § 27]. In this case we have the following result.

THEOREM 5. *Suppose $A(z)$ and $z^s B(1/z)$ have greatest common divisor*

$$D(z) = (z - \zeta_1) \cdots (z - \zeta_m) \qquad (m \geqq 1), \tag{55}$$

and let

$$\frac{A(z)}{D(z)} = A_1(z) = \alpha_0 + \cdots + \alpha_{r-m} z^{r-m}, \tag{56}$$

$$\frac{z^s B(1/z)}{D(z)} = z^{s-m} B_1(1/z) = \beta_{s-m} + \cdots \beta_0 z^{s-m}. \tag{57}$$

Then the characteristic polynomial $p_n(\lambda)$ in (47) *is given by*

$$p_n(\lambda) = \frac{(-1)^{m(k+1)+(r-1)(n-m)} c_r^n}{R_1[\zeta_1 \cdots \zeta_m]^s} \lambda^m \tilde{\Delta}_n(\lambda), \qquad n > k, \tag{58}$$

where

$$\tilde{\Delta}_n(\lambda) = |A_1(z), \cdots, z^{s-1} A_1(z), z^{n+s-m} B_1(1/z), \cdots, z^{n+k-m-1} B_1(1/z)|(\lambda), \tag{59}$$

and R_1 is the (*nonzero*) *value of the $k \times k$ determinant with rows $i = 1, \cdots, k$ as follows*:

(a) *For $1 \leqq i \leqq s$ there are $i-1$ zeros*; *then $\alpha_0, \cdots, \alpha_{r-m}$, then $s - i + m$ zeros.*

(b) *For $s + 1 \leqq i \leqq k$ there are $m + i - s - 1$ zeros*; *then $\beta_{s-m}, \cdots, \beta_0$, then $k - i$ zeros.*

Proof. Again we consider only ordinary points λ of $P(z; \lambda)$. For $1 \leqq j \leqq k$, $D(z_j)$ is a common factor of the jth column of the determinant in the numerator of $\Delta_n(\lambda)$. (See (46) and recall (40).) Removing these common factors shows that

$$\Delta_n(\lambda) = D(z_1) \cdots D(z_k) \tilde{\Delta}_n(\lambda), \tag{60}$$

with $\tilde{\Delta}_n(\lambda)$ as in (59), because of (56) and (57). From (55),

$$D(z_1) \cdots D(z_k) = \prod_{l=1}^{m} (z_1 - \zeta_l) \cdots (z_k - \zeta_l). \tag{61}$$

Since $z_1, \cdots, z_k$ are the zeros of $P(z; \lambda)$, (12) implies that

$$(z_1 - \zeta_l) \cdots (z_k - \zeta_l) = (-1)^k c_r^{-1} P(\zeta_l; \lambda), \qquad 1 \leqq l \leqq m.$$

But $A(\zeta_l) = 0$, so (6) and (12) imply that $P(\zeta_l; \lambda) = -\lambda \zeta_l^s$. This and (61) imply that

$$D(z_1) \cdots D(z_k) = (-1)^{m(k+1)} c_r^{-m} (\zeta_1 \cdots \zeta_m)^s \lambda^m.$$

This and (60) yield

$$(62)\qquad \Delta_n(\lambda)=(-1)^{m(k+1)}c_r^{-m}(\zeta_1\cdots\zeta_m)^s\lambda^m\tilde{\Delta}_n(\lambda).$$

An induction argument like the one used to prove (49) shows that

$$(63)\qquad \tilde{\Delta}_n(\lambda)=(-1)^{(r-1)(n-m)}R_1c_r^{-n+m}\lambda^{n-m}+\tilde{g}_n(\lambda),\qquad n\geqq m,$$

where $\deg\tilde{g}_n(\lambda)<n-m$. To see that $R_1\neq 0$, we observe that since $A_1(z)$ and $z^{s-m}B_1(1/z)$ are relatively prime and $A_1(0)\neq 0$, it follows that $A_1(z)$ and $z^sB_1(1/z)$ are relatively prime. Therefore, an argument like the one used in Theorem 2 to prove that $R\neq 0$ applies here.

From (62) and (63), the polynomial on the right of (58) is monic and of exact degree n. An argument similar to that used in the proof of Theorem 2 now establishes (58). This completes the proof of Theorem 4.

Laplace's development provides a convenient method for expanding the determinants in (46) and (59); see [11, § 5].

Now let $E_n(\lambda)$ be the solution space of the system

$$H_nX=\lambda X.$$

The following lemma is analogous to a lemma obtained in [11] for Toeplitz band matrices.

LEMMA 6. *Let λ and $z_1,\cdots,z_q$ be as in Definition 1. Then λ is an eigenvalue of H_n if and only if there are polynomials*

$$(64)\qquad f(z)=C_0+\cdots+C_{s-1}z^{s-1},\qquad g(z)=D_0+\cdots+D_{r-1}z^{r-1},$$

such that the polynomial

$$(65)\qquad h(z)=f(z)A(z)+z^{n+s}g(z)B(1/z)$$

is not identically zero and has zeros at $z_1,\cdots,z_q$ with multiplicities at least $m_1,\cdots,m_q$; i.e.,

$$(66)\qquad h^{(l)}(z_j)=0,\quad 0\leqq l\leqq m_j-1,\quad 1\leqq j\leqq q.$$

Moreover, if $S_n(\lambda)$ is the vector space of polynomials h of the form (64) *and* (65) *which satisfy* (66), *then*

$$\dim S_n(\lambda)=\dim E_n(\lambda).$$

Proof. A polynomial h of the stated form satisfies (66) if and only if the vector

$$Y=\mathrm{col}\,[C_0,\cdots,C_{s-1},D_0,\cdots,D_{r-1}]$$

satisfies the system

$$\Omega_n'Y=0.$$

Therefore, $\dim S_n(\lambda)=$ nullity of $\Omega_n'=$ nullity of $\Omega_n=\dim E_n(\lambda)$. (See the proof of Theorem 2.)

Lemma 6 implies the next two theorems. Since the proofs of these theorems are the same as those of Theorems 3 and 4 of [11], we omit them.

THEOREM 6. *If λ is an eigenvalue of H_n then*

$$\dim E_n(\lambda)\leqq\min\,(r,s).$$

THEOREM 7. *Suppose λ is an eigenvalue of H_n and* $\dim E_n(\lambda) \geqq 2$. *Then λ is also an eigenvalue of H_{n-1} (if $n > k+1$) and H_{n+1}; moreover,*

$$\dim E_{n-1}(\lambda) \geqq -1 + \dim E_n(\lambda)$$

and

$$\dim E_{n+1}(\lambda) \geqq -1 + \dim E_n(\lambda).$$

Acknowledgments. The author thanks T. N. E. Greville for reviving his interest in the matrices discussed here several years after the publication of [10]. The present paper was strongly influenced by the joint paper [7], and by Greville's subsequent work [2]-[6] on these matrices and on their applications to data smoothing. In particular, the crucial step of introducing the extrapolated components of U in Lemmas 4 and 5 was motivated by a similar device used by Greville in [4], [5], and [6] to extend symmetric smoothing formulas to the extremities of data.

REFERENCES

[1] D. BINI AND M. CAPOVANI, *Spectral and computational properties of band symmetric Toeplitz matrices*, Lin. Alg. Appl., 52/53 (1983), pp. 99-126.

[2] T. N. E. GREVILLE, *On a problem concerning band matrices with Toeplitz inverses*, Proc. 8th Manitoba Conference on Numerical Methods of Computation, Utilitas, Winnipeg, 1978, pp. 275-283.

[3] ———, *Bounds for the eigenvalues of Hermitian Trench matrices*, Proc. 9th Manitoba Conference Numerical Methods of Computation, Utilitas, Winnipeg, 1980, pp. 241-256.

[4] ———, *Moving-weighted-average smoothing extended to the extremities of the data.* I. *Theory*, Scand. Actuar. J. (1981), pp. 39-55.

[5] ———, *Moving-weighted-average smoothing extended to the extremities of the data.* II. *Methods*, Scand. Actuar. J. (1981), pp. 65-81.

[6] ———, *Moving-weighted-average smoothing extended to the extremities of the data.* III. *Stability and optimal properties*, J. Approx. Th., 33 (1981), pp. 43-58.

[7] T. N. E. GREVILLE AND W. F. TRENCH, *Band matrices with Toeplitz inverses*, Lin. Alg. Appl., 27 (1979), pp. 199-209.

[8] F. A. GRUNBAUM, *Toeplitz matrices commuting with tridiagonal matrices*, Lin. Alg. Appl., 40 (1981), pp. 25-36.

[9] ———, *Eigenvectors of a Toeplitz matrix: discrete version of prolate spheroidal wave functions*, this Journal, 2 (1981), pp. 136-141.

[10] W. F. TRENCH, *Weighting coefficients for the prediction of stationary time series from the finite past*, SIAM J. Appl. Math., 15 (1967), pp. 1502-1510.

[11] ———, *On the eigenvalue problem for Toeplitz band matrices*, Lin. Alg. Appl., 64 (1985), pp. 199-214.

[12] B. L. VAN DER WAERDEN, *Modern Algebra*, Frederick Ungar, New York, 1949.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

002

LU-DECOMPOSITIONS OF TRIDIAGONAL IRREDUCIBLE *H*-MATRICES*

W. J. HARROD†

Abstract. In this paper bounds are developed and investigated on the growth factors and the multipliers resulting from Gaussian elimination applied to an irreducible tridiagonal H-matrix. These results extend the study of the stability of Gaussian elimination without pivoting on certain tridiagonal matrices by Gunzburger and Nicolaides.

AMS(MOS) subject classification. 15A23

1. Background and information. By an *M-matrix* we will mean an $n \times n$ real matrix $A=(a_{ij})$ such that $a_{ij} \leqq 0$ for all $i \neq j$ and the principal minors of A are nonnegative (see Berman and Plemmons [1979, Chap. 6]). Assume that A is an irreducible tridiagonal matrix of order n, written in the form:

$$A=\begin{bmatrix} a_1 & b_1 & & 0 \\ c_1 & \ddots & \ddots & \\ & \ddots & \ddots & b_{n-1} \\ 0 & & c_{n-1} & a_n \end{bmatrix}.$$

Let σ_k denote the kth leading principal minor of A. Then it follows that A is an M-matrix if and only if

$$\begin{aligned} &c_j,\ b_j \leqq 0, && 1 \leqq j \leqq n-1,\\ &\sigma_j > 0, && 1 \leqq j \leqq n-1,\\ &\sigma_n \geqq 0. \end{aligned}$$

Research that involves concepts related to tridiagonal matrices is of widespread interest. Recently, papers have been published that investigate the stability of Gaussian elimination applied to tridiagonal Toeplitz matrices (Gunzburger and Nicolaides [1982]), characterizations of tridiagonal D-stable matrices (Carlson, et al. [1982], Carlson [1984]), and the characterization of nonnegative nonsingular tridiagonal matrices that belong to the class of inverse M-matrices (Inman [1983]).

If A is an irreducible matrix, then $c_j b_j \neq 0$, for $1 \leqq j \leqq n-1$. For notational purposes A will be denoted by

$$(1) \qquad A=[c_j,\ a_j,\ b_j].$$

If any of the sequences $\{c_j\}$, $\{a_j\}$, or $\{b_j\}$ are constant, then the corresponding unsubscripted variable will replace the subscripted variable in (1). A matrix $B=(b_{ij})$ is a Toeplitz matrix if there exists a sequence $\{\beta_k\}_{k=-n+1}^{n-1}$, such that $b_{ij}=\beta_{j-i}$, for $1 \leqq i, j \leqq n$. Hence, a tridiagonal Toeplitz matrix will be denoted by $A=[c, a, b]$.

Let A be a complex matrix of order n, such that A has all nonzero diagonal elements. We define its *comparison matrix* $\mathcal{M}(A)=(\delta_{ij})$ by

$$\delta_{ij}=\begin{cases} |a_{jj}| & \text{if } i=j,\\ -|a_{ij}| & \text{if } i \neq j. \end{cases}$$

* Received by the editors May 15, 1984 and in revised form September 15, 1984.

† Departments of Mathematics and Computer Science, University of Kentucky, Lexington, Kentucky 40506.

Then A is called an *H-matrix*, if $\mathcal{M}(A)$ is an M-matrix. Assume that $M=(m_{ij})$ is an M-matrix. We define the set Ω_M of complex matrices by

$$\Omega_M=\{A\in\mathbb{C}^{n\times n}|M\leqq\mathcal{M}(A)\}.$$

Thus $A=(a_{ij})\in\Omega_M$ if and only if

$$\begin{aligned}|a_{ij}|&\leqq|m_{ij}| &&\text{if } i\neq j,\\ |a_{jj}|&\geqq m_{jj} &&\text{if } i=j.\end{aligned}$$

We will define an *LU-decomposition* of a matrix A to be a decomposition $A=LU$, where L is a unit lower triangular matrix and U is an upper triangular matrix. Funderlic and Plemmons [1982] show that if there exists a vector x such that $x^TM\geqq 0$ and $x\gg 0$, then for each $A\in\Omega_M$ and each permutation matrix P, PAP^T has an LU-decomposition. Also, if A and M have the LU-decompositions

$$A=LU,\qquad M=L'U',$$

then

$$\begin{aligned}|l_{ij}|&\leqq|l'_{ij}| &&\text{for all } i \text{ and } j,\\ |u_{ij}|&\leqq|u'_{ij}| &&\text{for all } i\neq j,\\ |u_{jj}|&\geqq u'_{jj} &&\text{for all } j.\end{aligned}\tag{2}$$

Let $A^{(0)}=A$ and $A^{(k)}$ denote the matrix that results after the first $k-1$ steps of Gaussian elimination without pivoting applied to A. The rounding error analysis for Gaussian elimination without pivoting shows that the method is stable only when the entries in the matrix L and the matrices $A^{(k)}$ for $1\leqq k\leqq n-1$ do not grow excessively during the course of the elimination process. The *growth factor* g_A for Gaussian elimination applied to A is defined by

$$g_A=\max_{i,j,k}|a_{ij}^{(k)}|/\max_{i,j}|a_{ij}|.$$

In Wilkinson [1961], it was shown that if A is a strictly column diagonally dominant matrix, then for each $1\leqq i,j\leqq n$, $|l_{ij}|\leqq 1$ and $g_A\leqq 2$. Therefore Gaussian elimination, when applied to a column diagonally dominant tridiagonal matrix, is a stable method.

The purpose of this paper is to investigate the set of irreducible tridiagonal H-matrices with respect to the stability of the LU-decomposition without pivoting. We will show that pivoting is not necessary to control the growth in the entries of the matrices L and U. Explicit upper bounds will be provided for the values $|l_{ij}|$, $|u_{ij}|$ for $1\leqq i,j\leqq n$ and g_A.

2. Main results. Theorem 1 implies that an irreducible H-matrix admits an LU-decomposition.

THEOREM 1 (Fiedler and Pták [1962]). *Let A be an irreducible H-matrix. Then there exists a vector x such that $x^T\mathcal{M}(A)\geqq 0$ and $x\gg 0$. Moreover, if A is singular then $x^T\mathcal{M}(A)=0$.*

If A is an H-matrix, then $A\in\Omega_{\mathcal{M}(A)}$. Thus, if A is also irreducible, then A admits an LU-decomposition. In particular, an irreducible tridiagonal H-matrix $A=[c_j,a_j,b_j]$ admits an LU-decomposition, and it can be shown that

$$L=[m_j,1,0]\quad\text{and}\quad U=[0,\alpha_j,b_j],$$

where the multipliers m_j and the diagonal entries α_j satisfy the following difference equations

$$(3) \qquad \alpha_j = a_j - b_{j-1}m_{j-1}, \qquad 2 \leqq j \leqq n,$$

$$(4) \qquad m_j = c_j / \alpha_j, \qquad 1 \leqq j \leqq n-1,$$

and $\alpha_1 = a_1$.

Applying the above difference equations, some useful properties concerning the LU-decompositions of tridiagonal matrices are provided.

PROPOSITION 1. *Let $A = [c_j, a_j, b_j]$ be a tridiagonal H-matrix of order n, such that $c_j b_j \neq 0$ for $1 \leqq j \leqq n-1$. If $A = LU$ and $A^T = \bar{L}\bar{U}$, where*

$$L = [m_j, 1, 0], \qquad U = [0, \alpha_j, b_j],$$

and

$$\bar{L} = [\bar{m}_j, 1, 0], \qquad \bar{U} = [0, \bar{\alpha}_j, c_j],$$

then

$$(5) \qquad \begin{aligned} \alpha_j &= \bar{\alpha}_j, & 1 \leqq j \leqq n, \\ m_j &= \left(\frac{c_j}{b_j}\right) \bar{m}_j, & 1 \leqq j \leqq n-1. \end{aligned}$$

Proof. Combining (3) and (4), it follows that

$$\alpha_j = a_j - \frac{b_{j-1}c_{j-1}}{\alpha_{j-1}}, \qquad 2 \leqq j \leqq n.$$

Also, it can be shown that

$$\bar{\alpha}_j = a_j - \frac{b_{j-1}c_{j-1}}{\bar{\alpha}_{j-1}}, \qquad 2 \leqq j \leqq n.$$

Clearly, $\alpha_1 = \bar{\alpha}_1 = a_1$, so both α_j and $\bar{\alpha}_j$ satisfy the same difference equation, with the same initial condition. Therefore it follows that $\alpha_j = \bar{\alpha}_j$ for $1 \leqq j \leqq n$. Since $\bar{m}_j = b_j / \bar{\alpha}_j$ for $1 \leqq j \leqq n-1$ it can be shown that (5) follows. □

PROPOSITION 2. *Let $A = [c_j, a_j, b_j]$ be a tridiagonal M-matrix such that $c_j b_j \neq 0$ for $1 \leqq j \leqq n-1$. If $A = LU$ is the LU-decomposition of A and $U = (u_{ij})$, then*

$$|u_{ij}| \leqq \max_j \{a_j, |b_j|\}.$$

Proof. If $A = LU$ is the LU-decomposition of A, then L and U are both M-matrices. Thus it follows that $m_j \leqq 0$ and $\alpha_j > 0$ for $1 \leqq j \leqq n-1$ and $\alpha_n \geqq 0$. Since

$$\alpha_j = a_j - m_{j-1}b_{j-1}, \qquad 2 \leqq j \leqq n,$$

it follows that $\alpha_j \leqq a_j$ for $1 \leqq j \leqq n$ and the proof is complete. □

Therefore, when Gaussian elimination is applied to a tridiagonal irreducible M-matrix there is no growth in the magnitude of the elements of the matrix U, hence $g_A = 1$. Let σ_j denote the jth principal leading minor of $A = [c_j, a_j, b_j]$. Then, it can be shown, by induction, that the σ_j's satisfy the following difference equation

$$(6) \qquad \sigma_{j+1} - a_{j+1}\sigma_j + c_j b_j \sigma_{j-1} = 0, \qquad 1 \leqq j \leqq n-1$$

where $\sigma_1 = a_1$ and $\sigma_0 = 1$. Lemma 1 will be used to prove necessary and sufficient conditions for a Toeplitz tridiagonal irreducible matrix to be an M-matrix.

LEMMA 1. *Assume that $\alpha > 0$ and $\gamma > 0$. Then the difference equation*

$$(7) \qquad \sigma_{j+1} - \alpha\sigma_j + \gamma\sigma_{j-1} = 0$$

where $\sigma_1 = \alpha$ and $\sigma_0 = 1$, has positive solutions σ_j for $1 \leqq k \leqq n$ if and only if one of the following conditions is true:

1) $\alpha^2 - 4\gamma \geqq 0$,

2) $\alpha^2 - 4\gamma < 0$ *and* $\tan^{-1}\left(\frac{\sqrt{4\gamma - \alpha^2}}{\alpha}\right) < \frac{\pi}{n+1}$.

Proof. If $\alpha^2 - 4\gamma > 0$, then the solution to the difference equation (7) is

$$\sigma_j = \frac{\lambda_1^{j+1} - \lambda_2^{j+1}}{\lambda_1 - \lambda_2},$$

where $\lambda_{1,2} = (\alpha \pm \sqrt{\alpha^2 - 4\gamma})/2$. Clearly, it follows that $\lambda_1 > \lambda_2$. Thus, $\sigma_j > 0$ for all $1 \leqq j \leqq n$.

If $\alpha^2 - 4\gamma = 0$, then the solution to the difference equation (7) is

$$\sigma_j = (j+1)\left(\frac{\alpha}{2}\right)^j.$$

Thus, $\sigma_j > 0$ for all $1 \leqq j \leqq n$.

If $\alpha^2 - 4\gamma < 0$, then the solutions to the difference equation (7) is

$$\sigma_j = \gamma^{k/2}\frac{\sin((j+1)\theta)}{\sin(\theta)},$$

where $\theta = \tan^{-1}((\sqrt{4\gamma - \alpha^2})/\alpha)$. Since $(\sqrt{4\gamma - \alpha^2})/\alpha > 0$, it follows that $\theta \in (0, \pi/2)$. Thus, $\sin((j+1)\theta) > 0$ for $1 \leqq k \leqq n$ if and only if $0 < (j+1)\theta < \pi$ for $1 \leqq k \leqq n$. Therefore, it holds that

$$0 < \theta < \frac{\pi}{n+1}.$$

Hence, $\sigma_j > 0$ for all $1 \leqq j \leqq n$ if and only if

$$\tan^{-1}\left(\frac{\sqrt{4\gamma - \alpha^2}}{\alpha}\right) < \frac{\pi}{n+1}. \qquad \square$$

Next we state Theorem 2 which establishes necessary and sufficient conditions for A to be a Toeplitz tridiagonal M-matrix.

THEOREM 2. *Let $A = [c, \dot{a}, b]$ be a Toeplitz tridiagonal matrix of order n. If $a > 0$ and $b, c < 0$, then A is a nonsingular M-matrix if and only if one of the following conditions is true*:

i) $a^2 - 4bc \geqq 0$,

ii) $a^2 - 4bc < 0$ *and* $\tan^{-1}\left(\frac{\sqrt{4bc - a^2}}{a}\right) < \frac{\pi}{n+1}$.

Proof. If σ_j denotes the jth leading principal, the difference equation for σ_j reduces to the following:

$$\sigma_{j+1} - \sigma_j a + \sigma_{j-1} cb = 0, \qquad 1 \leqq j \leqq n-1.$$

Let $\alpha = a$ and $\gamma = bc$. Then it follows, by applying Lemma 1, that $\sigma_j > 0$ for $1 \leqq j \leqq n$ if and only if one of the conditions i) or ii) hold true. □

The matrix $A = [c, a, b]$, where $a > 0$ and $b, c < 0$, is a singular M-matrix if and only if $\sigma_j > 0$ for $1 \leqq j \leqq n-1$ and $\sigma_n = 0$. Let $\alpha = a$ and $\gamma = bc$. Then by examining the proof of Lemma 1, it is clear that for $\sigma_n = 0$, it must hold that $a^2 - 4bc < 0$. Corollary

1 provides necessary and sufficient conditions for A to be a singular irreducible M-matrix.

COROLLARY 1. *Let $A=[c, a, b]$ be a Toeplitz tridiagonal matrix of order n. If $a>0$ and $b, c<0$, then A is a singular M-matrix if and only if $a^2-4bc<0$ and*

$$\tan\left(\frac{\sqrt{4bc-a^2}}{a}\right)=\frac{\pi}{n+1}.$$

Theorem 3 provides a sufficient condition for a tridiagonal matrix to be a nonsingular M-matrix.

THEOREM 3. *Let $A=[c_j, a_j, b_j]$ be a tridiagonal matrix of order n, such that $a_j>0$ for $j=1, 2, \cdots, n$ and $b_j, c_j<0$ for $j=1, 2, \cdots, n-1$. Let $\alpha=\min_j a_j$ and $\gamma=(\max_j c_j)(\max_j b_j)$. If one of the following conditions is true, then A is a nonsingular M-matrix*:

i) $\alpha^2-4\gamma\geqq 0,$

ii) $\alpha^2-4\gamma<0$ *and* $\tan^{-1}\left(\frac{\sqrt{4\gamma-\alpha^2}}{\alpha}\right)<\frac{\pi}{n+1}.$

Proof. Let $c=-\max_j|c_j|$ and $b=-\max_j|b_j|$. Then $M=[c, \alpha, b]$ is a nonsingular M-matrix, by Lemma 1. Clearly $M\leqq A$. Thus it follows from well known properties of M-matrices (see Berman and Plemmons [1979, Chap. 6]) that A is a nonsingular M-matrix.

The solution of the homogeneous linear system $Ax=0$ where A is a singular irreducible M-matrix is important in many areas of mathematical sciences such as input-output analysis in economics (e.g., Berman and Plemmons [1979, Chap. 9], compartmental analysis tracer models (e.g., Funderlic and Mankin [1981]), and finite Markov chains (e.g., Berman and Plemmons [1979, Chap. 8]).

Theorem 4 provides a difference equation for the entries in a vector x such that $x^TA\geqq 0$ and $x\gg 0$, when A is a tridiagonal irreducible M-matrix.

THEOREM 4. *Let $A=[c_j, a_j, b_j]$ be a tridiagonal M-matrix of order n, such that $c_jb_j\neq 0$ for $1\leqq j\leqq n-1$. If $x=(x_j)$ where the x_j's satisfy the following difference equation*:

$$(8)\qquad \begin{aligned} &c_{j+1}x_{j+2}+a_{j+1}x_{j+1}+b_jx_j=0, \qquad 0\leqq j\leqq n-2,\\ &x_1=1, \qquad x_0=0, \end{aligned}$$

then $x\gg 0$ and $x^TA\geqq 0$. Furthermore, if A is a singular matrix, then $x^TA=0$.

Proof. Multiplying the difference equation (6) for the leading principal minors of A by $-(\prod_{l=1}^j|c_l|)^{-1}$ and since $|c_l|=-c_l$, it follows that

$$\left(\prod_{l=1}^{j+1}|c_l|\right)^{-1}\sigma_{j+1}c_{j+1}+\left(\prod_{l=1}^{j}|c_l|\right)^{-1}\sigma_j a_{j+1}+\left(\prod_{l=1}^{j-1}|c_l|\right)^{-1}\sigma_{j-1}b_j=0.$$

Let $x_{j+1}=(\prod_{l=1}^j|c_l|)^{-1}\sigma_j$ for $0\leqq j\leqq n$ and $x_0=0$. Then, the sequence $\{x_j\}_{j=0}^{n+1}$ satisfies the following difference equation:

$$c_{j+2}x_{j+1}+a_{j+1}x_{j+1}+b_jx_j=0, \qquad x_1=1, \quad x_0=0.$$

Therefore, $x_j>0$ for $1\leqq j\leqq n$ and $x_{n+1}\geqq 0$.

Let $z^T=x^TA$. Then it follows that

$$(9)\qquad z_j=\begin{cases} x_2c_1+x_1a_1 & \text{if } j=1,\\ x_{j+1}c_j+x_ja_j+x_{j-1}b_{j-1} & \text{if } 2\leqq j\leqq n-1,\\ x_na_n+x_{n-1}b_{n-1} & \text{if } j=n. \end{cases}$$

Applying the recurrence equation (8), it follows that $z_j = 0$ for $1 \leqq j \leqq n-1$. Therefore, $z^T = x^T A = \alpha e_n^T$ and it can be shown that $\alpha \geqq 0$, since A is an irreducible M-matrix. Therefore, $x^T A \geqq 0$. If A is a singular irreducible M-matrix, then A is almost monotone (see Fiedler and Pták [1962]); that is, $x^T A \geqq 0$ implies that $Ax = 0$. □

Next, a strict upper bound on the moduli of the elements of L will be derived, where $A = LU$ is the LU-decomposition of a tridiagonal irreducible H-matrix. Lemma 2 provides a useful expression for the multipliers of A.

LEMMA 2. *Let $A = [c_j, a_j, b_j]$ be a tridiagonal H-matrix of order n, such that $b_j c_j \neq 0$ for $1 \leqq j \leqq n-1$. If $A = LU$ is the LU-decomposition of A and $L = [m_j, 1, 0]$, then*

$$m_j = c_j \frac{\sigma_{j-1}}{\sigma_j} \quad \text{for } j = 1, 2, \cdots, n-1,$$

where σ_j denotes the jth principal leading minor of A and $\sigma_0 = 1$.

THEOREM 5. *Let $A = [c_j, a_j, b_j]$ be an M-matrix of order n, such that $c_j b_j \neq 0$ for $1 \leqq j \leqq n-1$. Let $M_A = \max_j |m_j|$, where m_j is the jth multiplier of A. Then*

$$|m_j| \leqq \frac{a_{j+1}}{|b_j|}, \qquad 1 \leqq j \leqq n-1,$$

and thus

$$M_A \leqq \max_j \frac{a_{j+1}}{|b_j|}.$$

Proof. Solving the difference equation (6) for σ_{j+1} yields

$$\sigma_{j+1} = a_{j+1}\sigma_j - c_j b_j \sigma_{j-1}.$$

Since $\sigma_{j+1} > 0$ for $1 \leqq j \leqq n-2$ and $\sigma_n \geqq 0$, it follows that

$$0 \leqq a_{j+1}\sigma_j - c_j b_j \sigma_{j-1}, \qquad 1 \leqq j \leqq n-1.$$

Thus

$$|c_j| \cdot \frac{\sigma_{j-1}}{\sigma_j} \leqq \frac{a_{j+1}}{|b_j|}, \qquad 1 \leqq j \leqq n-1.$$

Applying Lemma 2, the proof is complete. □

COROLLARY 2. *Assume that $A = [c, a, b]$ is a Toeplitz tridiagonal M-matrix of order $n \geqq 3$, such that $cb \neq 0$. Then, either $M_A < 2$ or $M_{A^T} < 2$.*

Proof. It follows by Theorem 4 that $M_A \leqq a/|b|$ and that $M_{A^T} \leqq a/|c|$. If $|b| \geqq a$ and $|c| \geqq a$ then $cb \geqq a^2$, or $0 \geqq a^2 - bc$. This is a contradiction since $\sigma_2 = a^2 - bc$ and $\sigma_2 > 0$. Thus, either $|b| < a$ or $|c| < a$.

Assume that $|b| < a$ and $|c| \geqq a$. Then $a/|c| \leqq 1$ and $M_{A^T} \leqq 1$. Assume that $|b| \geqq a$ and $|c| < a$. Then $a/|b| \leqq 1$ and $M_A \leqq 1$. Assume that $|b| < a$ and $|c| < a$. If A is column diagonally dominant, then it is well known that $M_A \leqq 1$. Thus assume that A is not column diagonally dominant. Therefore it must be true that $a < |b| + |c|$. If $|c| < |b|$, then $a < 2|b|$ and $M_A < 2$. If $|b| \leqq |c|$, then $a < 2|c|$ and $M_{A^T} < 2$. □

In a recent paper by Gunzburger and Nicolaides [1982], it is shown that it is possible to find a Toeplitz tridiagonal matrix $A = [c, a, b]$, where $b = c = -1$ and $0 < a < 2$, such that the multipliers may become excessively large. However, A is not an M-matrix.

Assume that $A = [c_j, a_j, b_j] \in \mathbb{C}^{n \times n}$ is an irreducible tridiagonal H-matrix. Then $\mathcal{M}(A) = [-|c_j|, |a_j|, -|b_j|]$ is an irreducible tridiagonal M-matrix. Also, $\mathcal{M}(A)$ admits an LU-decomposition $\mathcal{M}(A)=L'U'$, where

$$\begin{aligned} &L'=[m_j', 1, 0], \qquad U'=[0, \alpha_j', -|b_j|], \\ &m_j'=-|c_j|/\alpha_j', \qquad 1\leqq j\leqq n-1, \end{aligned} \tag{10}$$

$$\begin{aligned} &\alpha_j'=|a_j|+m_{j-1}'|b_j|, \qquad 2\leqq j\leqq n, \\ &\alpha_1'=|a_1|. \end{aligned} \tag{11}$$

Theorem 6 provides upper bounds for M_A and g_A, when A is an irreducible H-matrix.

THEOREM 6. *Assume that $A=[c_j, a_j, b_j]$ is a tridiagonal H-matrix of order n, such that $c_jb_j\neq 0$ for $j=1,2,\cdots,n-1$. If $A=LU$ where*

$$\begin{aligned} &L=[m_j, 1, 0], \qquad U=[0, \alpha_j, b_j], \\ &m_j=c_j/\alpha_j, \qquad 1\leqq j\leqq n-1, \\ &\alpha_j=a_j-m_{j-1}b_{j-1}, \qquad 2\leqq j\leqq n, \\ &\alpha_1=a_1, \end{aligned} \tag{12}$$

then

$$\begin{aligned} &|m_j|\leqq\frac{|a_{j+1}|}{|b_j|}, \qquad 1\leqq j\leqq n-1, \\ &|\alpha_j|\leqq 2|a_j|, \qquad 1\leqq j\leqq n, \end{aligned}$$

so that $M_A\leqq\max_j(|a_{j+1}|)/(|b_j|)$, and $g_A\leqq 2$.

Proof. Applying Theorem 5 to the matrix $\mathcal{M}(A)$, it follows that

$$|m_j'|\leqq\frac{|a_{j+1}|}{|b_j|}, \qquad 1\leqq j\leqq n-1,$$

where m_j' is given by (10). Since $A\in\Omega_{\mathcal{M}(A)}$, we can apply (2) and get

$$|m_j|\leqq|m_j'|\leqq\frac{|a_{j+1}|}{|b_j|}, \qquad 1\leqq j\leqq n-1. \tag{13}$$

Therefore

$$M_A\leqq\max_j(|a_{j+1}|/|b_j|).$$

Taking the absolute value of both sides of (12), it follows that

$$|\alpha_j|\leqq|a_j|+|b_{j-1}|\,|m_{j-1}|, \qquad 1\leqq j\leqq n.$$

If we apply (13), one obtains

$$|\alpha_j|\leqq|a_j|+|b_{j-1}|\frac{|a_j|}{|b_{j-1}|}, \qquad 1\leqq j\leqq n,$$

or

$$|\alpha_j|\leqq 2|a_j|, \qquad 1\leqq j\leqq n.$$

Therefore $g_A\leqq 2$. □

Thus, for a tridiagonal H-matrix A the growth of the entries in the matrix U is bounded above by two and the multipliers m_j cannot become excessively large.

Finally, we can derive a lower bound for the multipliers of $A=[c_j, a_j, b_j]$. Applying Theorem 6, it follows that

$$\frac{1}{2|a_j|} \leqq \frac{1}{|\alpha_j|}, \qquad 1 \leqq j \leqq n. \tag{14}$$

From (4) we have that $|m_j| = |c_j|/|\alpha_j|$ for $j = 1, 2, \cdots, n-1$. Thus, by Theorem 6 and (14)

$$\frac{|c_j|}{2|a_j|} \leqq |m_j| \leqq \frac{|a_{j+1}|}{|b_j|}, \qquad 1 \leqq j \leqq n-1.$$

When A is a reducible tridiagonal M-matrix it is possible that one of the multipliers m_j may become excessively large. As an illustration, consider the following nonsingular reducible M-matrix

$$A = \begin{pmatrix} 2 & -2 & 0 \\ \varepsilon-2 & 2 & 0 \\ 0 & -1 & 3 \end{pmatrix}, \qquad 1 > \varepsilon > 0,$$

which admits the following LU-decomposition:

$$A = \begin{pmatrix} 1 & 0 & 0 \\ (\varepsilon-2)/2 & 1 & 0 \\ 0 & 1/\varepsilon & 1 \end{pmatrix} \begin{pmatrix} 2 & -2 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & -3 \end{pmatrix}.$$

Then, $M_A = 1/\varepsilon$ so that M_A approaches infinity as ε approaches zero.

REFERENCES

A. BERMAN AND R. PLEMMONS (1979), *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York.

D. CARLSON (1984), *Controllability, inertia, and stability for tridiagonal matrices*, Linear Algebra Appl., 56, pp. 207-220.

D. CARLSON, B. DATTA AND C. JOHNSON (1982), *A semi-definite Lyapunov theorem and the characterization of tridiagonal D-stable matrices*, this Journal, 3, pp. 293-304.

M. FIEDLER AND V. PTÁK (1962), *On matrices with nonpositive off-diagonal elements and positive principal minors*, Czechoslovak Math. J., 12, pp. 382-400.

R. FUNDERLIC AND J. MANKIN (1981), *Solution of homogeneous systems of linear equations arising from compartmental models*, SIAM J. Sci. Statis. Comput., 2, pp. 315-383.

R. FUNDERLIC, M. NEUMANN AND R. PLEMMONS (1982), *LU-decompositions of generalized diagonally dominant matrices*, Numer. Math., 40, pp. 57-69.

R. FUNDERLIC AND R. PLEMMONS (1982), *LU-decomposition of M-matrices by elimination without pivoting*, Linear Algebra Appl., 41, pp. 99-110.

M. GUNZBURGER AND R. NICOLAIDES (1982), *Stability of gaussian elimination without pivoting on tridiagonal toeplitz matrices*, Linear Algebra Appl., 45, pp. 21-28.

I. INMAN (1983), *Tridiagonal and upper triangular inverse M-matrices*, Linear Algebra Appl., 55, pp. 93-104.

J. WILKINSON (1961), *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8, pp. 281-330.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

003

INVERSE PROBLEMS FOR MEANS OF MATRICES*

WILLIAM N. ANDERSON, JR.† AND GEORGE E. TRAPP‡

Abstract. Given two positive semidefinite Hermitian matrices A and B there are natural definitions for their arithmetic and harmonic means. In this work we consider the following question: Given positive semidefinite matrices C and D, when do there exist positive semidefinite matrices A and B such that C is the arithmetic mean of A and B and D is the harmonic mean of A and B. Uniqueness questions are also answered. Similar questions are answered concerning the geometric mean.

AMS(MOS) subject classifications. 15A24, 15A45

1. Introduction. Many of the familiar means of nonnegative scalars may be defined on pairs of positive Hermitian operators. The reader is referred to [12] for an introduction to means of matrices; also see the paper by Kubo and Ando [10]. The primary object of this paper is to determine when two positive semidefinite matrices C and D can be the arithmetic and harmonic, or arithmetic and geometric, means of matrices A and B. These questions were considered in the scalar case by Gauss, see [8]. In particular Gauss derived the scalar equation of Theorem 2 below.

We restrict our attention to the finite dimensional case. A positive semidefinite matrix will always be assumed to be Hermitian. We will use the partial order induced by positive semidefiniteness, that is $A \geqq B$ means $A-B$ is Hermitian positive semidefinite (HSD).

For HSD matrices A and B, the *arithmetic mean* of A and B is defined by

$$A \bigtriangleup B = (A+B)/2.$$

If A and B are invertible, the *parallel sum* of A and B, denoted $A:B$, is defined by

$$A:B=(A^{-1}+B^{-1})^{-1}.$$

In the general case, one may define the parallel sum using Lemma 1e below, or by the formula $A:B=A(A+B)^{-}B$, where any 1− inverse may be chosen (that is $(A+B)(A+B)^{-}(A+B)=A+B$).

Using the parallel sum, the *harmonic mean* of A and B is defined by

$$A\,!\,B=2(A:B).$$

We next wish to consider a generalization of the geometric mean of scalars. In order to define the *geometric mean* of the HSD matrices A and B, let $A_0=A$ and $B_0=B$. Inductively, let $A_{n+1}=A_n \bigtriangleup B_n$, and $B_{n+1}=A_n\,!\,B_n$. The sequences A_n and B_n both converge to the geometric mean, which we denote by $A \# B$ [3], [10]. If appropriate invertibility is available, the following definitions are equivalent to the iterative process [10], [7] and [3].

$$A \# B = A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{1/2},$$

$$A \# B = (AB^{-1})^{1/2}B,$$

$$A \# B = (A+B)((A+B)^{-1}A(A+B)^{-1}B)^{1/2}.$$

* Received by the editors March 15, 1985, and in revised form April 2, 1985. Portions of this paper were presented at the 1980 IEEE International Symposium on Circuits and Systems, Houston, TX, April 1980.

† Department of Mathematics and Computer Science, Fairleigh Dickinson University, Teaneck, New Jersey 07666.

‡ Department of Statistics and Computer Science, West Virginia University, Morgantown, West Virginia 26506.

If A and B commute, it is easy to see that $A \# B = (AB)^{1/2}$; it is this property and Lemma 1 below which justify calling $A \# B$ the geometric mean.

It is immediate from the iterative definition that $A \# B = (A \triangle B) \# (A\,!\,B)$.

The properties of the means which we shall need are summarized in Lemma 1. Proofs are contained in references [1] and [10].

LEMMA 1. *Let A and B be* HSD *matrices, and let C be a matrix. Let m denote any of the means arithmetic, geometric or harmonic. Then*

(a) $AmA = A$.

(b) $AmB = BmA$.

(c) $C(AmB)C^* \leqq (CAC^*)m(CBC)$ *with equality if C is invertible.*

(d) *If* $B_1 \leqq B_2$ *then* $AmB_1 \leqq AmB_2$.

(e) *If A_n and B_n are sequences of matrices monotonically decreasing to A and B respectively, then $A_n m B_n$ decreases monotonically to AmB.*

(f) $A\,!\,B \leqq A \# B \leqq A \triangle B$.

(g) range $(A\,!\,B) =$ range $(A) \cap$ range (B).

(h) range $(A \# B) =$ range $(A) \cap$ range (B).

In the next section, we consider the inverse mean problems; in § 3, we discuss various duality questions pertaining to the inverse means.

2. Inverse mean problems. In this section, we consider two inverse mean problems. First we are interested in when can two HSD matrices C and D be the arithmetic and harmonic means respectively of two other HSD Matrices A and B? Our Theorem 1 below supplies the existence and uniqueness results. Theorem 2 answers a similar question for the arithmetic and geometric means.

THEOREM 1. *Let C and D be* HSD *matrices, with $C \geqq D$. Then there exist* HSD *matrices A and B such that*

$$C = A \triangle B, \tag{1a}$$

$$D = A\,!\,B. \tag{1b}$$

Moreover, if we require $A \geqq B$, then the solution is unique.

Proof. In the scalar case we can solve for A and B using the quadratic formula, obtaining $A = C \pm (C^2 - CD)^{1/2}$. We are thus led to consider the solutions

$$A = C + C \# (C - D), \tag{2a}$$

$$B = C - C \# (C - D). \tag{2b}$$

First we note A defined by (2a) is HSD, since $C - D$ is HSD by hypothesis. To see that B, defined by (2b) is HSD, note that $C \geqq C - D$, so that $C = C \# C \geqq C \# (C - D)$.

It is clear that $A + B = 2C$, so that (1a) is satisfied.

In order to show that (1b) holds, let us first assume that C is invertible. Let E denote $C^{-1/2}DC^{-1/2}$. Then the following sequence of equalities holds:

$$\begin{aligned} A\,!\,B &= (C + C \# (C - D))!(C - C \# (C - D)) \\ &= C^{1/2}((I + I \# (I - E))!(I - I \# (I - E)))C^{1/2} \\ &= C^{1/2}((I + (I - E)^{1/2})!(I - (I - E)^{1/2}))C^{1/2} \\ &= 2C^{1/2}((I + (I - E)^{1/2}):(I - (I - E)^{1/2}))C \end{aligned}$$

$$\begin{aligned}&=2C^{1/2}((I+(I-E)^{1/2})(2I)^{-1}(I-(I-E)^{1/2}))C^{1/2}\\&=C^{1/2}(I-(I-E))C^{1/2}\\&=C^{1/2}EC^{1/2}=D.\end{aligned}$$

We therefore have the formula

$$D=(C+C\#(C-D))!(C-C\#(C-D))$$

for invertible C, and the general result holds by taking limits, using Lemma 1(e). Therefore formulas (2) furnish a solution to equations (1). It remains to consider uniqueness.

Again let us assume that C is invertible. Then from equations (1) we pre-multiply and post-multiply by $C^{-1/2}$ to obtain the equivalent system

(3a) $$I=X+Y,$$

(3b) $$Z=X:Y$$

where $Z=4C^{-1/2}DC^{-1/2}$. Note, we already know that (3) has at least one solution.

For any solution pair X, Y we must satisfy the following equation, where we replace Y by $I-X$ from (3a)

(4) $$Z=X:(I-X)=X-X^2.$$

We have assumed that $A\geqq B$ and therefore $X\geqq Y$. Since $Y=I-X$, the condition $X\geqq Y$ is equivalent to the condition $X\geqq I/2$. Using equation (4), we obtain the following equation

$$(X-I/2)^2=I/4-Z.$$

Since the HSD square root of a HSD matrix is unique, we have that Z uniquely determines $X-I/2$ or equivalently X is unique.

If C is not invertible, we note that from (1a) we must have range (A) ⊂ range (C) and range (B) ⊂ range (C). Then from (1b) we have range (D) ⊂ range (C). Since all matrices under consideration have their ranges contained in the range of C, we may transform from (1) to (3) using the Moore-Penrose generalized inverse and the result follows.

Finally, we note that in the absence of the condition $A\geqq B$ the solution will not be unique unless $C=D$.

THEOREM 2. *Let C and D be HSD matrices with $C\geqq D$. Then there exist HSD matrices A and B such that*

(5a) $$C=A\triangle B,$$

(5b) $$D=A\# B.$$

Moreover, if we require $A\geqq B$, then the solution is unique.

Proof. As before, we consider the scalar case to motivate the solutions

(6a) $$A=C+(C+D)\#(C-D),$$

(6b) $$B=C-(C+D)\#(C-D).$$

Since $C\geqq D$ it is clear that A defined by (6a) is HSD. To see that B defined by (6b) is HSD, we note the following equality and inequality of Lemma 1f

$$C=(C+D)/2+(C-D)/2\geqq(C+D)\#(C-D).$$

It is clear that $A+B=2C$, so that (5a) holds. For (5b), we again first assume that C is invertible. Then, letting $E=C^{-1/2}DC^{-1/2}$, we have the following string of equalities

$$\begin{aligned}A \# B &= (C+(C+D)\#(C-D))\#(C-(C+D)\#(C-D))\\ &= C^{1/2}((I+(I+E)\#(I-E))\#(I-(I+E)\#(I-E)))C^{1/2}\\ &= C^{1/2}((I+(I-E^2)^{1/2})\#(I-(I-E^2)^{1/2}))C^{1/2}\\ &= C^{1/2}(I-(I-E^2))^{1/2}C^{1/2}\\ &= C^{1/2}EC^{1/2}=D.\end{aligned}$$

The noninvertible case is similar to the proof of Theorem 1. The uniqueness argument is analogous to that given in the proof of Theorem 1, except that (4) is replaced by

$$Z=(X(1-X))^{1/2}.$$

3. Duality. An important concept in electrical network theory is duality. The parallel sum of matrices, derived from the parallel connection of networks, is the natural dual of the ordinary sum of matrices (the series connection). Duality concepts may also be extended to means of matrices.

For every mean there is a *dual* mean, see [10]. The dual of the mean m is denoted by $m^{\perp}$ and defined by $Am^{\perp}B=(B^{-1}mA^{-1})^{-1}$ if the indicated inverses exist. Noninvertible cases are defined using Lemma 1(e).

The duality of ordinary and parallel addition, which is apparent from the definition of the parallel sum, yields that the arithmetic and harmonic means are duals [1], [6]. The geometric mean is self-dual, [10]. The dual of the difference $A-B$ is the parallel difference $A\div B$, which is defined by $(A^{-1}-B^{-1})^{-1}$ when the indicated inverses exist. In general, if $A\leqq B$ and range (B) = range $(B-A)$ then $A\div B$ may be defined by a limiting argument, or by the explicit formula $A\div B=A(B-A)^{+}B$, where the Moore-Penrose generalized inverse is used, see [2] or [11]. The dual to the equation $A=B+(A-B)$ is then the equation $A=B:(A\div B)$.

The dual of Theorem 1 merely interchanges equations (1a) and (1b), yielding nothing new. The duals to formulas (2a) and (2b) give alternate expressions for the solutions (1). The dual to Theorem 2 is more complicated and deserves a separate treatment. In Theorem 3 below, we consider the inverse mean problem for the geometric and harmonic means.

THEOREM 3. *Let C and D be* HSD *matrices with $C\leqq D$, and* range (C) = range (D). *Then there exist positive semidefinite matrices A and B with*

(7a) $$C=A\,!\,B,$$

(7b) $$D=A\#B.$$

Moreover, if C and D are invertible, the condition $A\leqq B$ will ensure uniqueness of the solution.

Proof. First let us consider the invertible case. From $C\leqq D$ we have $C^{-1}\geqq D^{-1}$ so that $DC^{-1}D\geqq DD^{-1}D=D\geqq C$. Then the hypotheses of Theorem 1 hold so that there exist A and B with $A\triangle B=DC^{-1}D$ and $A\,!\,B=C$. For this A and B we then have the following equalities

$$\begin{aligned}A\#B&=(A\triangle B)\#(A\,!\,B)\\ &=C\#(DC^{-1}D)\end{aligned}$$

$$
\begin{aligned}
&= C^{1/2}(C^{1/2}DC^{-1}DC^{-1/2})^{1/2}C^{1/2}\\
&= C^{1/2}(C^{-1/2}DC^{-1/2}C^{-1/2}DC^{-1/2})^{1/2}C^{1/2}\\
&= C^{1/2}C^{-1/2}DC^{-1/2}C^{1/2}\\
&= D.
\end{aligned}
$$

Thus A and B furnish a solution for equations (7).

For uniqueness, we reverse the above computation and observe that if A and B satisfy (7), then $A \triangle B = DC^{-1}D$. Theorem 1 then implies that the solution to (7) is uniquely determined by the condition $A \leqq B$.

If C and D are not invertible, they still must have the same range. The above existence argument holds with the inverses replaced by the Moore-Penrose generalized inverses. The resulting solutions A and B will be zero on the orthogonal complement of the range of C and D. Either (but not both) of A or B may be prescribed arbitrarily on this orthogonal complement without affecting (7), so that uniqueness does not hold in this case.

When the parallel difference $C \div D$ is defined, the solutions to (7) may be written as the duals to (6), that is

$$A = C:((C:D) \# (C \div D)), \tag{8a}$$

$$B = C \div ((C:D) \# (C \div D)). \tag{8b}$$

Formulas (8) may be verified by taking the dual of every line in the proof of Theorem 2.

Acknowledgment. The authors wish to thank the referees for improving the clarity of the proof of Theorem 1.

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., R. J. DUFFIN AND G. E. TRAPP, *Parallel subtraction of matrices*, Proc. Nat. Acad. USA, 69 (1972), pp. 2530-2531.

[3] W. N. ANDERSON, JR., T. D. MORLEY AND G. E. TRAPP, *A characterization of parallel subtraction*, Proc. Nat. Acad. Sci. USA, 76 (1979), pp. 3599-3601.

[4] T. ANDO, *Concavity of certain maps on positive definite matrices and applications to Hadamard products*, Lin. Alg. Appl., 26 (1979), pp., 203-241.

[5] ——, *On the arithmetic-geometric-harmonic mean inequalities for positive definite matrices*, Lin. Alg. Appl., 52 (1983), pp. 31-37.

[6] K. V. BHAGWAT AND R. SUBRAMANIAN, *Inequalities between means of positive operators*, Math. Proc. Camb. Phil. Soc., 83 (1978), pp. 393-401.

[7] H. J. CARLIN AND D. NOBLE, *Circuit properties of couple dispersive lines with applications to waveguide modeling*, in Network and Signaling Theory, A NATO Advanced Study Institute, Peregrinus, London, 1972, pp. 258-269.

[8] K. F. GAUSS, *Werke*, 3, B. G. Teubner, Leipzig, 1917, pp. 361-365.

[9] P. R. HALMOS, *Finite Dimensional Vector Spaces*, Van Nostrand, New York, 1958.

[10] F. KUBO AND T. ANDO, *Means of positive linear operators*, Math. Ann., 246 (1980), pp. 205-224.

[11] E. L. PEKAREV AND Y. L. SMULYAN, *Parallel addition and parallel subtraction of operators*, Izv. Akad. Nauk. USSR, 40 (1976), pp. 366, 387; Math. Izv, 10 (1976), pp. 351-370.

[12] G. E. TRAPP, *Hermitian semidefinite matrix means and related matrix inequalities—An introduction*, Lin Multilin. Alg., 16 (1984), pp. 112-123.

[13] W. PUSZ AND L. WORNOWICZ, *Functional calculus for sesquilinear maps and the purification map*, Rep. Math. Phys., 9 (1975), pp. 159-170.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

004

INCOMPLETE FACTORIZATION OF SINGULAR *M*-MATRICES*

J. J. BUONI†

Abstract. In 1981, Varga and Cai characterized those M-matrices A (perhaps singular) which admit a factorization into M-matrices L and U ($A = LU$) where L is required to be a nonsingular and lower triangular M-matrix and U is required to be an upper triangular M-matrix, a result that was first proved by Fiedler and Ptak (1962) in the case when A is nonsingular. Because this factorization may, as a result of fill-in, produce a lower triangular matrix which is considerably less sparse than A, one attempts to control the fill-in of the factorization of A by means of a graph. This method leads to the concept of incomplete factorizations of A. Meijerink and van der Vorst (1977)·who have shown that incomplete factorizations of nonsingular M-matrices are possible, while Manteuffel (1980) has extended this result to the H-matrix case. The purpose of this paper is to give a condition on a singular M-matrix which guarantees the incomplete factorization of a singular M-matrix.

Key words. M-matrices, incomplete factorization

AMS(MOS) subject classifications. 65F05, 65F10

1. Introduction. An $n \times n$ M-matrix $A = (a_{ij})$ is said to *admit an LU factorization into n × n M-matrices* if A can be expressed as

$$A = LU \tag{1.1}$$

where $L := (1_{ij})$ is an $n \times n$ *lower triangular* M-matrix (i.e. $1_{ii} > 0$, $1_{ij} < 0$ for all $i > j$ and $1_{ij} = 0$ for all $j > i$, where $1 \leqq i, j \leqq n$) and where $U := (u_{ij})$ is an $n \times n$ *upper triangular* M-matrix (i.e., $u_{ii} > 0$, $u_{ij} < 0$ for all $i < j$ and $u_{ij} = 0$ for all $j < i$, where $1 \leqq i, j \leqq n$). A well-known result of Fiedler and Ptak in 1962 (cf. [4]) gives that any nonsingular M-matrix admits such an LU factorization (1.1) into M-matrices, with L and U both nonsingular. In 1977, Kuo [6] extended this result by showing that any $n \times n$ irreducible M-matrix (singular or not) admits an LU factorization (1.1) into M-matrices, with say, L nonsingular. In 1981, Varga and Cai (cf. [13, Thm. 1]) characterized those M-matrices which admit an LU factorization into M-matrices with L nonsingular with the following result:

THEOREM 1. *Let A be an $n \times n$ M-matrix. Then the following are equivalent*:

1. *A admits an LU factorization into M-matrices with nonsingular L.*

2. *For every proper subset $s := (s_1, \cdots, s_k)$ of $\langle n \rangle := (1, \cdots, n)$ for which $A\langle s \rangle$ is singular and irreducible, there is no path in the directed graph $G_n(A)$ of A from vertex v_t to vertex v_g for any $t > s_k$ and q in s.*

Because the factorization in (1.1) may, as a result of fill-in, produce a lower triangular matrix L which is considerably less sparse than A, one attempts to control the fill-in of the factorization of A by means of a graph, an idea which seems to have first been suggested by Varga [10] as a specific technique for generating regular splittings (cf. [11, p. 88]) of certain finite difference equations. This method leads to incomplete factorizations of A, and is described below.

For n any positive integer, let $A = (a_{ij})$ with real entries, and let G (for graph) denote any nonempty set of ordered pairs of integers (i, j), with $i \leqq i, j \leqq n$ and with $i \neq j$. Then given any $A = (a_{ij})$ and given any graph G, we attempt to produce a splitting of A

$$A = M - N, \tag{1.2}$$

* Received by the editors January 7, 1985, and in revised form April 24, 1985.

† Department of Mathematical and Computer Sciences, Youngstown State University, Youngstown, Ohio 44555.

where $M=LU$ and L and U are sparse lower and upper triangular nonsingular M-matrices with the properties:

$$l_{ij}=0 \quad \text{if } (i,j) \text{ in } G,$$
$$u_{ij}=0 \quad \text{if } (i,j) \text{ in } G.$$

Meijerink and van der Vorst [8] have shown that (1.2) is possible when A is an nonsingular M-matrix, while Manteuffel [7] extended this result to the H-matrix case.

The purpose of this paper is to show that a condition which is slightly weaker than that of Theorem 1 will guarantee the incomplete factorization of a singular M-matrix into the form of (1.2).

2. Main results. In this section we establish our main results.

LEMMA 1. *Let A be an M-matrix. Let the elements of $B=(b_{ij})$ satisfy the relations*

$$a_{ij} \leqq b_{ij} \leqq 0 \quad \textit{for } i \neq j, \qquad 0 \leqq a_{ii} \leqq b_{ii}. \tag{2.1}$$

Then

1. *B is an M-matrix.*
2. *If B is a singular M-matrix, then A is a singular M-matrix.*
3. *If A is an irreducible M-matrix and if $b_{ij}=0$ for some $i \neq j$ where $a_{ij} \neq 0$ then B is nonsingular.*

Proof. For A an M-matrix, 1) and 2) follow immediately from arguments which appear in [13], while 3) may be found in [5]. □

Set $P=((i,j)$: $a(i,j)=0$ where $i \neq j)$, i.e. the graph of the off-diagonal zero entries of A.

The *Incomplete Factorization Algorithm* (IFA) may be given recursively as follows:

$$\begin{aligned} A^0 &= A, \\ C_k &= A^{k-1}+R^k, \\ A^k &= L_k C_k \end{aligned} \tag{2.2}$$

where $k>0$,

$$r^k_{ij}=\begin{cases} -a^{k-1}_{ij} & \text{if } (i,j) \text{ in } P, \\ 0 & \text{otherwise,} \end{cases}$$

and L_k is equal to the unit matrix, except for the kth column, which written row-wise is as follows:

$$[0,0,\cdots 1, -c^k_{k+1,k}/c^k_{kk}, \cdots, -c^k_{nk}/c^k_{kk}] \quad \text{when } c^k_{kk} \neq 0; \tag{2.3}$$

otherwise,

L_k is the identity matrix.

Remark. The algorithm fails if $c^{k-1}_{kk}=0$ and $c^{k-1}_{tk} \neq 0$ for $t>k$.

LEMMA 2. *Let $i,j \geqq k$ and $i \neq j$. Then $\widehat{i,j}$ is an arc in the graph of A iff $\widehat{i,j}$ is an arc in C_k.*

Proof. Without loss of generality, we may assume that $k=2$. If $\widehat{i,j}$ is an arc in the graph of A, then $a_{ij} \neq 0$, then from (2.2), i.e. first step in the Gaussian elimination,

$$a^1_{ij}=-a^0_{i1}*a^0_{1j}/a^0_{11}+a^0_{ij} \quad \text{for } i,j>1, \qquad a^0_{11} \neq 0, \tag{2.4}$$

and $r^2_{ij}=0$ because $\widehat{i,j}$ is not in $P(a_{ij}$ is not zero).

Since $a^0_{tq}<0$ for $t \neq q$, then $a^1_{ij}<0$, i.e. $\widehat{i,j}$ is an arc in the graph in A^1. Now $r^2_{ij}=0$ implies that $\widehat{i,j}$ is an arc in C_2.

Conversely, let $\widehat{i,j}$ be an arc in C_2. Since R^2 removes arcs from A^1 which were in P, then $\widehat{i,j}$ is not an element of P, e.g. originally it was not removed from A. Hence $\widehat{i,j}$ is an arc in A.

The case for $a_{11}^0 = 0$ follows immediately. □

If A is a reducible M-matrix, then there exists a permutation matrix P and an integer s with $1 \leqq s \leqq n$ such that A may be placed in a normal reduced form

$$(2.5) \qquad PAP^t = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1s} \\ 0 & A_{22} & \cdots & A_{2s} \\ 0 & 0 & \cdots & A_{3s} \\ \vdots & \vdots & & \vdots \\ \cdot & \cdot & & A_{ss} \end{bmatrix}$$

where each A_{ij} is an irreducible M-matrix. Furthermore, if A is singular then A_{jj} is singular for some j [13].

LEMMA 3. *Let A an M-matrix. If at any completed step k of* (2.2), *a_{kk}^{k-1} in A^k is zero then A is singular.*

Proof. Since a_{kk}^{k-1} is a diagonal element of the upper triangular block of A^k, then A^k and C_k, are singular and L_k is constructed nonsingular. By the second part of Lemma 2, one finds that A^{k-1} is singular. Backtracking yields that $A = A^0$ is singular. □

THEOREM 4. *Let A an M-matrix, and let $s = (s_1, \cdots, s_k)$ be any subset of $\langle n \rangle :=$ $(1, \cdots, n)$ for which $A\langle s \rangle$ is singular and irreducible. If*

$$(2.6) \qquad a_{tp} = 0 \quad \textit{for all } t > s_k \textit{ and all } p \textit{ in } s$$

then the Incomplete Factorization Algorithm (2.2) *does not fail at any step.*

Proof. Assume (2.2) fails at step k, i.e. one cannot form $A_k = L_k C_k$. Then it follows from the remark following (2.2) that $c_{kk}^k = 0$ and $c_{rk}^k \neq 0$ for some $r > k$. However, the algorithm does work through $k-1$ steps and (2.2) did work for $A[\langle k-1 \rangle]$. But as $c_{kk}^k = 0$, then from (2.2) $a_{kk}^{k-1} = 0$, and when one views $A^{k-1}[\langle k \rangle]$ one obtains

$$A^{k-1}[\langle k \rangle] = \begin{bmatrix} A^{k-1}[\langle k-1 \rangle] & * \\ **** & c_{kk} \end{bmatrix}$$

i.e. $A^{k-1}[\langle k \rangle]$ is singular; hence, a singular M-matrix. Therefore, similar to Lemma 3, $A[\langle k \rangle]$ is a singular M-matrix by Lemma 1.

Now as in [13], we set $1 \leqq s_1 < s_2 < \cdots < s_j = k$, be the largest subset of $\langle k \rangle$ for which $A[s]$ is irreducible. Since $s_j = k$ and $a_{kk}^{k-1} = 0$, then $A^{k-1}[s]$ is singular and by Lemma 1, $A[s]$ is singular. Now since $c_{rk} \neq 0$, it follows from Lemma 2 that $a_{rk} \neq 0$, which contradicts (2.6). □

3. Regular splittings. For a real $n \times n$ matrix A (possibly singular), the splitting $A = M - N$ is regular iff $M^{-1} \geqq 0$ and $N \geqq 0$. Unfortunately, for singular matrices A, regular splittings may only semi-converge, (cf. [1, p. 154]).

THEOREM 5. *If A is an $n \times n$ irreducible M-matrix which satisfies* (2.6), *then the splitting $A = LU - R$ produced in* (2.2) *is regular, provided R is a nonzero matrix.*

Proof. Theorem 4 produced candidates for L, U, and R. It remains to construct the splitting.

First observe as in [8] that

$$L_k R^m = R^m \quad \text{if } k < m, \qquad L_k R^k > 0.$$

Proceeding as in [8], one immediately obtains from (2.2) that

$$\begin{aligned} A^{n-1} &= L_{n-1}C_{n-1} = L_{n-1}A^{n-2} + L_{n-1}R^{n-1} \\ &= L_{n-1}L_{n-2}\cdots L_1A^0 + L_{n-1}\cdots L_1R^1 + L_{n-1}\cdots L_2R^1 + \cdots + L_{n-1}R^{n-1}. \end{aligned}$$

By considering these equations, we find

$$A^{n-1} = L_{n-1}\cdots L_1(A + R^1 + \cdots + R^{n-1}).$$

Let us now set $U = A^{n-1}$, $L = (L_{n-1}\cdots L_1)^{-1}$ and $R = R^1 + \cdots + R^{n-1}$; then

$$LU = A + R, \quad R \geqq 0.$$

It only remains to show that U is nonsingular and that $U^{-1} \geqq 0$. To this end, assume that A is irreducible. Then after the first elimination step, A^1 of (2.2) has a nonzero element in $(1, 1)$ position and an irreducible matrix in the $(2, \cdots, n)$ position (cf. [13, p. 187]). If R^2 deletes an element from this submatrix then by Lemma 1.c, we are finished. If not, then continue the argument until some R^k does delete a nonzero element from an irreducible submatrix of some A^{k-1}. □

4. Examples. In [13] the following example is given to illustrate an M-matrix which does not admit an LU factorization. However, by Theorem 4 it admits an incomplete factorization.

Example 1.

$$A = \begin{bmatrix} 6 & -1 & 0 & 0 & 0 & 0 \\ -1 & 6 & 0 & -1 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6 & -1 \\ -1 & 0 & 0 & 0 & -1 & 6 \end{bmatrix}, \quad L = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{6} & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ -\frac{1}{6} & 0 & 0 & 0 & -\frac{1}{6} & 1 \end{bmatrix},$$

$$U = \begin{bmatrix} 6 & -1 & 0 & 0 & 0 & 0 \\ 0 & \frac{35}{6} & 0 & -1 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6 & -1 \\ 0 & 0 & 0 & 0 & 0 & \frac{35}{6} \end{bmatrix}, \quad R = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{6} & 0 & 0 & 0 & 0 \end{bmatrix}.$$

It is clear that A does not satisfy the conditions of Theorem 1 (cf. [13]); however, $A = LU - R$.

In the following example we illustrate an example which does not satisfy the hypothesis of Theorem 6 and does not factor.

Example 2.

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

then in the notation of (2.2) one finds

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad L_1A = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 1 \end{bmatrix}.$$

Hence, the incomplete factorization fails.

For singular matrices regular splittings of the form $A = M - N$ may *only* converge for relaxation parameters c such that $0 < c < 1$, as substantiated by the following example due to Hans Schneider that appeared in [2].

Example 3.

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

Then

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -\frac{1}{2} & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{bmatrix}, \qquad U = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

and

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

Now M^{-1} may be written as

$$\begin{bmatrix} 2 & 1 & 1 & \frac{1}{2} & \frac{1}{2} \\ 1 & 1 & 1 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix},$$

while $M^{-1}N$ is

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}. \qquad \square$$

Acknowledgments. The author would like to gratefully acknowledge the remarks of R. Plemmons, R. S. Varga and the referee which led to improvements in the manuscript.

REFERENCES

[1] A Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[2] J. J. Buoni, M. Neumann and R. S. Varga, *Theorems of Stein-Rosenberg type. III. The singular case*, Lin. Alg. Appl., 42 (1982), pp. 183-198.

[3] Ky Fan, *Note on M-matrices*, Quart. J. Math. Oxford Ser. (2), 11 (1960), pp. 43-49.

[4] M. Fiedler and V. Ptak, *On matrices with nonpositive off-diagonal elements and positive principal minors*, Czech. Math. J., 12 (1962), pp. 382-400.

[5] R. E. Funderlic and R. J. Plemmons, *A combined direct-iterative method for certain M-matrix linear systems*, this Journal, 5 (1984), pp. 33-42.

[6] I-wen Kuo, *A note on factorization of singular M-matrices*, Lin. Alg. Appl., 16 (1977), pp. 217-220.

[7] T. Manteuffel, *An incomplete factorization technique for positive definite linear systems*, Math. Comp., 34 (1980), pp. 473-497.

[8] J. A. Meijerink and H. A. van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31 (1977), pp. 148-162.

[9] M. Neumann and R. Plemmons, *Convergent nonnegative matrices and iterative methods for consistent linear systems*, Numer. Math., 31 (1978), pp. 265-279.

[10] R. S. Varga, *Factorization and normalized iterative methods*, in Boundary Problems in Differential Equations, R. E. Langer, ed., Univ. Wisconsin Press, Madison, WI, 1960.

[11] ———, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[12] R. S. Varga, E. B. Saff and V. Mehrmann, *Incomplete factorization of matrices and connection with H-matrices*, SIAM J. Numer. Anal., 17 (1980), pp. 787-793.

[13] R. S. Varga and D.-Y. Cai, *On the LU factorization of M-matrices*, Numer. Math., 38 (1981), pp. 179-192.

[14] ———, *On the LU factorization of M-matrices: cardinality of the set $P(A)$*, this Journal, 3 (1982), pp. 250-259.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

005

PACKINGS BY COMPLETE BIPARTITE GRAPHS*

P. HELL† AND D. G. KIRKPATRICK‡

Abstract. Given any set $\mathcal{B}$ of complete bipartite graphs, we ask whether a graph H admits a $\mathcal{B}$-factor, i.e., a spanning subgraph, each of whose components is a member of $\mathcal{B}$. More generally, we seek in H a maximum $\mathcal{B}$-packing, i.e., a $\mathcal{B}$-factor of a maximum size subgraph of H. We first treat the interesting special case when $\mathcal{B}$ is a set of stars. The results are generalized to arbitrary $\mathcal{B}$ in the last section. We prove for most of these problems that they are $\mathcal{NP}$-hard; we also show that the remaining problems admit polynomial algorithms based on augmenting configurations. The simplicity of these algorithms, as well as the implied min-max theorems, resemble the theory of matchings in bipartite, rather than general, graphs.

AMS(MOS) subject classifications. 05C70, 68R10

1. Introduction.

Let $\mathcal{G}$ be a (possibly infinite) set of (finite) graphs. A *$\mathcal{G}$-factor* of a (finite) graph H is a spanning subgraph of H consisting of a number of disjoint copies of elements of $\mathcal{G}$. Equivalently, a $\mathcal{G}$-factor of H is the union G of subgraphs $G_1, G_2, \cdots, G_d$ of H whose vertex-sets partition $V(H)$ and such that each G_i is isomorphic to an element of $\mathcal{G}$. The *G-factor problem* can be described as follows:

INSTANCE: A graph H
QUESTION: Does H admit a $\mathcal{G}$-factor?

It is worth stressing that the set $\mathcal{G}$ is fixed, and not part of the instance. Also note that a $\{K_2\}$-factor of H is precisely a perfect matching of H, and hence we can view our problem as a kind of generalized matching problem.

As in the case of matchings, we may wish to find a $\mathcal{G}$-factor of a maximum-size subgraph of H. Formally, a *$\mathcal{G}$-packing* of a graph H is a $\mathcal{G}$-factor G of a subgraph H' of H; the vertices of H' are said to be *saturated*, the other vertices of H *exposed* by G. The number of vertices saturated by a $\mathcal{G}$-packing G is called the *size* of G, and a $\mathcal{G}$-packing of maximum size is called a *maximum* $\mathcal{G}$-packing.

Clearly, every H admits a $\{K_1\}$-factor. As noted above, the $\{K_2\}$-factor problem is precisely the perfect matching problem, and hence admits a polynomial-time solution [12]–[14]. Maximum matching algorithms can also be used to solve the $\{G\}$-factor and $\{G\}$-packing problems in the case when each component of G is K_1 or K_2 [28], [29]. We have shown that in all other cases the $\{G\}$-factor problem is $\mathcal{NP}$-complete, [28], [29]. We have also studied the $\mathcal{G}$-packing problems for various families $\mathcal{G}$, [20]–[23], [28], [29]; other recent works can also be viewed in this light [1]–[3], [8]–[10], [34]. In particular, for any family of complete graphs $\mathcal{G}$ there exist polynomial algorithms when K_1 or $K_2 \in \mathcal{G}$ [9], [20], [23], [29] and it can be shown that the $\mathcal{G}$-factor problem is $\mathcal{NP}$-hard in all other cases [20], [29].

In this paper we concentrate on the case where $\mathcal{G}$ is a family of complete bipartite graphs.

Some theoretical and practical motivation for the study of $\mathcal{G}$-factors and $\mathcal{G}$-packings was given in [28], [29]. In particular, packings by complete bipartite graphs

* Received by editors September 3, 1982, in revised form February 8, 1985.

† Departments of Computing Science and Mathematics, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6. The research of this author was supported by the Natural Sciences and Engineering Research Council of Canada under grant A5075.

‡ Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5. The research of this author was supported by the Natural Sciences and Engineering Research Council of Canada under grant A3583.

(and especially by stars) are of interest as a fruitful area in which analogues of results from traditional matching theory [4]-[6], [12]-[14], [26], [36], [37] are to be found.

2. Packings by sequential star sets. A *star*, S_i, is the complete bipartite graph $K_{1,i}$. A *sequential star set* $\mathcal{S}$ is either $\mathcal{S}=\{S_1, S_2, \cdots\}$ or $\mathcal{S}=\{S_1, S_2, \cdots, S_k\}$ for some positive integer k.

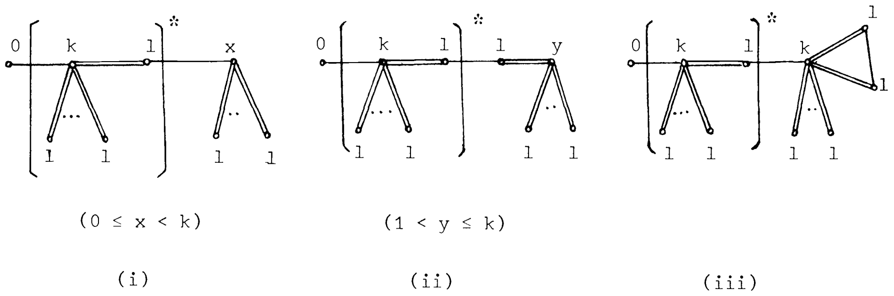
Our results on star packings date to the summer of 1979. They were discussed by the first author in several talks on generalized matchings and star packings, among others [19]; cf. also [29, Thm. 5.4] and [22]. We have since become aware of the related (and earlier) work in [7], [31]; moreover, other papers have since appeared [2], [3] which also directly or indirectly duplicate some of our results. Consequently we will restrict ourselves in this section to a brief outline of the main results; a more detailed version is available from the authors, [22].

PROPOSITION 1. *A graph H has an $\{S_1, S_2, \cdots\}$-factor if and only if it has no isolated vertices.*

Remark 1. Proposition 1 follows from the observation of Las Vergnas [31, Remark 3.5], that H has an $\{S_1, S_2, \cdots, S_k\}$-factor if and only if it has a $(1, k)$-factor, i.e., a spanning subgraph with all degrees between 1 and k. In fact, Las Vergnas' proof suggests an $O(|E|)$ algorithm to find a star-factor in a graph H without isolated vertices (clearly, a graph with isolated vertices has no star-factor): Examine each edge uv of H in turn, and delete it if both u and v have degree greater than 1 (updating the degrees of u and v if uv was deleted). After all edges have been examined, we are left with a star-factor G of H. Evidently, if all degrees of H were between 1 and k, then G will be an $\{S_1, S_2, \cdots S_k\}$-factor, so the same algorithm can be used to modify a $(1, k)$-factor of H to an $\{S_1, S_2, \cdots, S_k\}$-factor of H.

Remark 2. The algorithm discussed in Remark 1 will, when applied to an arbitrary graph H, find a maximum $\{S_1, S_2, \cdots\}$-packing; thus the maximum $\{S_1, S_2, \cdots\}$-packing problem is in $\mathcal{P}$. Moreover, the $\{S_1, S_2, \cdots, S_k\}$-factor problem can also be solved in polynomial time, because the $(1, k)$-factor problem can be solved in polynomial time [14], [35]. Finally, as we explain below, we shall give polynomial algorithms for (among others) the maximum $(1, k)$-packing problem (i.e., finding a subgraph of a given graph H with the maximum number of vertices and all degrees between 1 and k) [24], which can, using the algorithm in Remark 1, be translated to maximum $\{S_1, S_2, \cdots, S_k\}$-packing algorithms. A direct polynomial algorithm for the maximum $\{S_1, S_2, \cdots, S_k\}$-packing problem, based on familiar augmenting path techniques, is inherent in Theorems 1 and 2 below (cf. Remark 6).

Let G be an $\{S_1, S_2, \cdots, S_k\}$-packing of a graph H. Figure 1 sets out three basic augmenting configurations in H with respect to G. Figure 2 describes the corresponding augmentations. In these figures we depict the edges of G by double lines and the other



$(0 \leq x < k)$ $(1 < y \leq k)$

(i) (ii) (iii)

FIG. 1. *Basic augmenting configurations.*

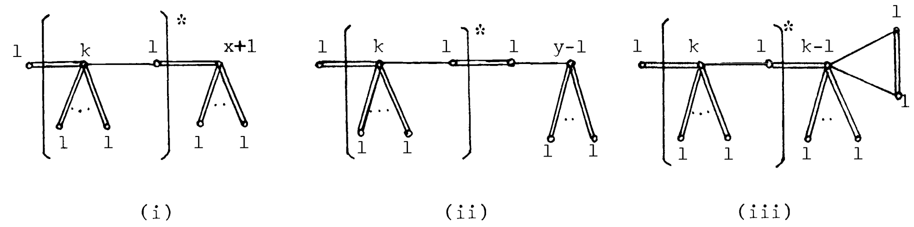


FIG. 2. *Corresponding augmented configurations.*

edges of H by single lines. Each vertex is labelled by its degree in G. The shorthand of Fig. 3a stands for any number ($\geqq 0$) of repetitions of Fig. 3b. Note that all vertices depicted (or implied by the shorthand) are assumed to be distinct.

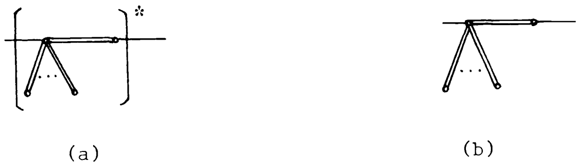


FIG. 3. *Shorthand for augmenting configurations.*

It should be clear from Figs. 1 and 2 that the absence of basic augmenting configurations is a necessary condition for G to be maximum. In fact, it is a sufficient condition as well:

THEOREM 1. *An $\{S_1, S_2, \cdots, S_k\}$-packing G of H is maximum if and only if it admits no basic augmenting configuration.*

Remark 3. It is useful to view entire augmenting configurations to see how a packing G is modified. However, as Figs. 1 and 2 make clear, augmentation takes place along a path in H. Indeed, we can characterize the presence of basic augmenting configurations in terms of familiar alternating paths. (An *alternating path* starts at an exposed vertex and alternates edges of $H-G$ and G). H admits a basic augmenting configuration with respect to G if and only if it admits an odd length alternating path ending at a vertex v whose degree d in G is less than k. We call such a path an *augmenting path* with respect to G. Note that each basic augmenting configuration contains an augmenting path.

On the other hand, if such a path exists then one of minimum length can be expanded to configuration (i) (if $d>1$) or (ii) (if $d=1$ and the star of G containing v is different from all the preceding stars on the path) or (iii) (if $d=1$ and the star containing v is the same as the star containing its predecessor on the path). (Because of minimality, it is easy to see that the star containing v cannot be the same as any other star meeting the path.)

Theorem 1 will now follow from the proof of the following Theorem 2.

THEOREM 2. *Let $k \geqq 2$. The size of a maximum $\{S_1, S_2, \cdots, S_k\}$-packing of H is equal to the minimum, over all $T \subseteq V(H)$, of $n+k\cdot|T|-i_T$ where $n=|V(H)|$ and i_T denotes the number of isolated vertices of $H-T$.*

Remark 4. Theorem 2 (although discovered independently) is a special case of [31, Prop. 4.1], if one takes into account the relation between $\{S_1, \cdots, S_k\}$-factors and $(1, k)$-factors discussed earlier (Remark 1). In Proposition 4.1, Las Vergnas gives the rank function of the matroid on $V(H)$ whose independent sets are vertex-sets $V' \subseteq V(H)$ which can be saturated by some $\{S_1, S_2, \cdots, S_k)$-packing of H; in that terminology, our Theorem 2 only gives the rank of the entire matroid. (However, the extension to the rank of any set is easy, cf. the proof of [31, Prop. 4.1].)

COROLLARY. *Let $k \geqq 2$. A graph H admits an $\{S_1, S_2, \cdots, S_k\}$-factor if and only if it does not admit a set T of vertices whose deletion results on more than $k|T|$ isolated vertices.*

The corollary was also discovered by [3], [7], [31] (and [2] when $k=2$). Our proof of Theorem 2 is different from that of Las Vergnas, but similar to the recent proof of the corollary by [3]. Hence we shall only give an outline:

Outline of proof of Theorem 2. If T is any subset of $V(H)$, then at least $i_T - k \cdot |T|$ vertices are exposed in any $\{S_1, S_2, \cdots, S_k\}$-packing G. Hence the maximum number of saturated vertices is at most $n-(i_T-k\cdot|T|)=n+k\cdot|T|-i_T$ for any $T \subseteq V(H)$. To prove the equality, we shall define an $\{S_1, S_2, \cdots, S_k\}$-packing G and a set T such that G saturates precisely $n+k\cdot|T|-i_T$ vertices of H. Let G be any $\{S_1, S_2, \cdots, S_k\}$-packing of H which admits no basic augmenting configuration. Let U denote the set of vertices exposed in G. Let T (respectively W) denote the set of vertices reachable from some vertex $u \in U$ by a nontrivial alternating path of odd (respectively even) length. Since G admits no augmenting configuration in H, it follows by Remark 3 that every vertex $t \in T$ has degree k in G. All vertices of $U \cup W$ are clearly isolated in $H-T$ and hence $i_T \geqq |U|+|W|$. Thus G saturates

$$n-|U| \geqq n+|W|-i_T=n+k\cdot|T|-i_T$$

vertices. □

It follows from the above proof that an $\{S_1, S_2, \cdots S_k\}$-packing G which does not admit a basic augmenting configuration admits a set $T \subseteq |V(H)|$ such that G saturates $n+k\cdot|T|-i_T$ vertices and hence G is a maximum packing. This completes the proof of Theorem 1 when $k \geqq 2$. When $k=1$, the basic configurations (ii) and (iii) are impossible and the statement reduces to the well-known augmenting path theorem [5].

There is an interesting distinction between the $\{S_1, S_2, \cdots, S_k)$-packing problem when $k=1$ and when $k \geqq 2$. While the former is exactly the matching problem in general graphs, the latter situation is similar to the theory of bipartite graphs, cf. Remarks 5 and 6.

Remark 5. Theorem 2 is false when $k=1$. For instance when $H=K_3$, the minimum of $n+|T|-i_T$ is 3. In fact, it follows from [4] (cf. also [15], [18], [37]) that the minimum of $n+|T|-i_T$ is the maximum size of a $\{K_2, C_3, C_5, C_7, \cdots\}$-packing. Note however that when H is bipartite such a packing is necessarily an $\{S_1\}$-packing, and so Theorem 2 does hold for bipartite graphs; it simply becomes a reformulation of König's theorem [30].

Remark 6. Theorem 1 and Remark 3 following it suggest an algorithm for maximum $\{S_1, S_2, \cdots, S_k\}$-packings: Begin with the empty packing G ($E(G)=\varnothing$), and given an $\{S_1, S_2, \cdots, S_k\}$-packing, find an augmenting path, expand it to an augmenting configuration, and update the $\{S_1, S_2, \cdots, S_k\}$-packing as explained after Theorem 1. In fact, when $k=1$, this is the typical approach to maximum matchings [5], [12], [13]. In the case of matchings, finding an augmenting path turns out to be at least conceptually easier in the case when H is bipartite: In general, a vertex can be reached from a fixed exposed vertex by augmenting paths of both even and odd lengths, and sometimes only paths of one parity can be continued; this leads to the need to blossom, [12], [13], in maximum matching algorithms based on the search for augmenting paths. In bipartite graphs, each vertex can be reached from a fixed vertex by either odd paths only or even paths only; therefore a breadth first search will be sufficient to identify an augmenting path. (Despite this intuitive simplification, the best current implementations of maximum matching algorithms have the same time bound $O(|E|\cdot\sqrt{|V|})$ for both bipartite, and general graphs [26], [33].) Similarly, for the $\{S_1, S_2, \cdots, S_k\}$-packing algorithms, when $k \geqq 2$, the vertices labeled k can only be reached by odd augmenting paths and those labeled 1 by even augmenting paths. Thus a breadth first search will find an augmenting path (in time $O(|E|)$) and the straightforward implementation of our maximum $\{S_1, S_2, \cdots, S_k\}$-packing algorithm would run in time $O(|E|\cdot|V|)$. Employing ideas similar to [11], [26], we can improve this time bound to $O(|E|\cdot\sqrt{|V|})$, [22]. We do not present this algorithm here, because in a companion paper [24] we plan to give a general algorithm to find a degree-constrained subgraph with maximum number of vertices, which has the same time bound of $O(|E|\cdot\sqrt{|V|})$ in the special case when each degree is constrained to be between 1 and k. (As pointed out in Remarks 1 and 2, such a degree-constrained subgraph can be modified to a maximum $\{S_1, S_2, \cdots, S_k\}$-packing in time $O(|E|)$.)

Remark 7. There is one more sense in which the result of Las Vergnas [31] is more general than Theorem 2. Namely, in his case, the stars used for packing in H are not all bound by the same number k, but the bound $f(v)$ depends on the vertex v: A *maximum starred* $(1, f)$*-packing* of H is a star packing of H of maximum size such that the star with center at v is $S_1, S_2, \cdots$, or $S_{f(v)}$. If each $f(v) \geqq 2$ both Theorem 1 and Theorem 2 remain valid, with obvious modification: In the augmenting configurations, replace each occurrence of k by the corresponding $f(v)$; in Theorem 2, the statement becomes:

The size of a maximum starred $(1, f)$-packing of H is equal to the minimum, over all $T \subseteq V(H)$ of

$$n + \sum_{t \in T} f(t) - i_T.$$

While this result also follows from [31], our proof based on augmenting configurations (as outlined above) offers an algorithm, as well as an interesting alternative to the proofs in [31]. (Cf. also [27] for the case of bipartite graphs, and for an interesting application.)

Remark 8. We note an important difference between the $(1, k)$-packing and the $\{S_1, \cdots, S_k\}$-packing problems. As noted above, they are equivalent when we maximize the number of vertices. However, they are substantially different when we maximize the number of their edges. A $(1, k)$-packing with maximum number of edges can be found in polynomial time, [14], [35]. The same problem is $\mathcal{NP}$-complete for $\{S_1, S_2, \cdots, S_k\}$-packings, since a graph H with n vertices has an $\{S_k\}$-factor if and only if the maximum number of edges in an $\{S_1, S_2, \cdots, S_k\}$-packing of H is $k/(k+1)\cdot n$. (The $\{S_k\}$-factor problem is $\mathcal{NP}$-complete for $k \geqq 2$, [28], [29].)

3. Packings by general star sets. Since describing a set of stars $\mathcal{S} \subseteq \{S_1, S_2, \cdots\}$ amounts to giving a set of positive integers $I_{\mathcal{S}} = \{i | S_i \in \mathcal{S}\}$, it should not be surprising that there exist star sets $\mathcal{S}$ for which the $\mathcal{S}$-factor problem is undecidable. This fact, as well as the fact that decidable $\mathcal{S}$-factor problems can be arbitrarily complex can be deduced from [25] and statement (a) of the following proposition:

PROPOSITION 2. (a) *The $\mathcal{S}$-factor problem is at least as hard as the membership problem for $I_{\mathcal{S}}$* (b) *If the membership problem for $I_{\mathcal{S}}$ is in $\mathcal{NP}$, then so is the $\mathcal{S}$-factor problem.*

Proof. Statement (a) follows from the observation that $i \in I_{\mathcal{S}}$ if and only if S_i has an $\mathcal{S}$-factor. Statement (b) is easy to deduce from the definition of $\mathcal{NP}$, [17]. □

The emphasis of our paper is on star sets $\mathcal{S}$ for which the membership problem for $I_{\mathcal{S}}$ is in $\mathcal{NP}$. Hence according to Proposition 2, all $\mathcal{NP}$-hardness results we state translate to $\mathcal{NP}$-completeness results for such star sets $\mathcal{S}$.

THEOREM 3. *The $\mathcal{S}$-factor problem is $\mathcal{NP}$-hard unless $\mathcal{S}$ is a sequential star set.*

Proof. The $\{S_2\}$-factor problem is $\mathcal{NP}$-complete by [17]. If $\mathcal{S}$ is not a sequential set and $\mathcal{S} \neq \{S_2\}$, then for some $t \geqq 3$ and $r < t$, $S_r \notin \mathcal{S}$, $S_{r+1} \in \mathcal{S}$, and $S_t \in \mathcal{S}$. (We may have $r+1=t$.) We shall reduce to the $\mathcal{S}$-factor problem the following well-known problem (t-dimensional matching):

INSTANCE: An integer n and a subset P of $\{1, 2, \cdots, n\}^t$

QUESTION: Does P admit a subset M with n elements no two of which agree in any coordinate? (Such a subset M is called a t-dimensional matching of P.)

This will prove Theorem 3, as the t-dimensional matching problem with $t \geqq 3$ is known to be $\mathcal{NP}$-complete, [17].

Suppose $P = \{\langle p_1^i, p_2^i, \cdots, p_t^i\rangle \colon 1 \leqq i \leqq l\}$ and consider the graph H_p defined as follows (see Fig. 4). The vertices of H_p are $A \cup B \cup C \cup D$, where $A = \{a_j^k \colon 1 \leqq k \leqq l-n, 1 \leqq j \leqq r\}$, $B = \{b_k \colon 1 \leqq k \leqq l-n\}$, $C = \{c_i \colon 1 \leqq i \leqq l\}$ and $D = \{1, 2, \cdots, n\} \times \{1, 2, \cdots, t\}$. In H_p each b_k is adjacent to all a_j^k and all c_i; moreover, each c_i is adjacent to $(p_1^i, 1), (p_2^i, 2), \cdots, (p_t^i, t)$.

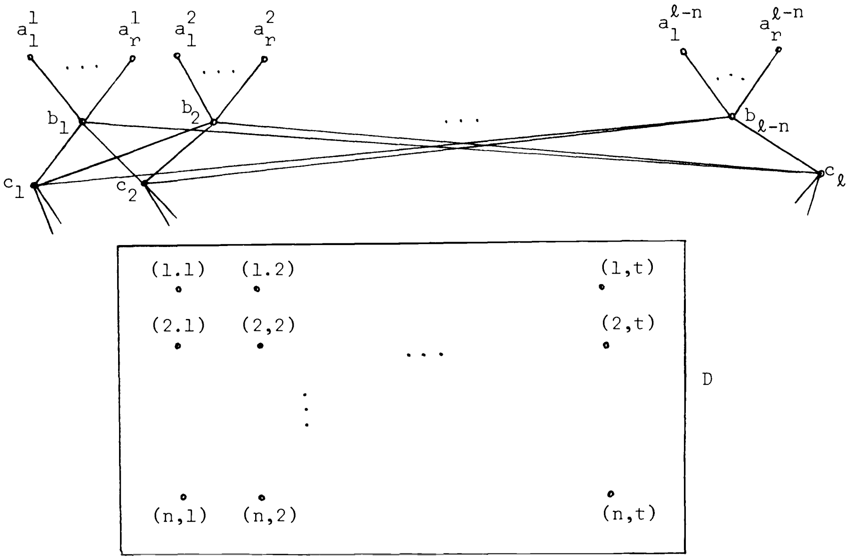


FIG. 4. *The graph H_p.*

H_p can clearly be constructed from P in polynomial time. We claim that H_p admits an $\mathcal{S}$-factor if and only if P admits a t-dimensional matching. If M is a t-dimensional matching of P, then each of the n t-tuples in M defines a copy of S_t with the center at some c_i and the endpoints in D. These stars contain n vertices c_i and all vertices of D. The remaining $l-n$ vertices c_i together with all vertices b_k and a_j^k have an obvious partition into copies of S_{r+1}; hence H_p has an $\mathcal{S}$-factor. Conversely, in any $\mathcal{S}$-factor H_p each vertex b_k lies on a star S_u with $u \geqq r+1$, and hence at most $l-(l-n)=n$ of the vertices c_i belong to stars involving D. Since each such star uses at most t vertices of D, there must be precisely n vertices c_i forming centers of stars S_t with all t other vertices in D. The t-tuples corresponding to these n stars define a t-dimensional matching $M \subseteq P$. □

4. The general $\mathcal{B}$-factor problem. Let $\mathcal{B}$ be an arbitrary set of complete bipartite graphs. In this section we show that the case where $\mathcal{B}$ is a sequential set of stars is essentially the only one for which we should expect a polynomial solution to the $\mathcal{B}$-factor problem; in all other cases the problem is $\mathcal{NP}$-hard.

The qualifier "essentially" can be made precise with the following notions. The set $\mathcal{B}$ is said to be *reducible* if any one of its elements, say B, admits a $\mathcal{B}-\{B\}$-factor; otherwise $\mathcal{B}$ is said to be *irreducible.* It should be clear that for every set $\mathcal{B}$ there is a unique irreducible subset $\mathcal{B}'\subseteq\mathcal{B}$ (the *kernel* of $\mathcal{B}$) such that for each $B\in\mathcal{B}$, B admits a $\mathcal{B}'$-factor. In the following lemma we observe that it is sufficient to focus our attention on irreducible sets of complete bipartite graphs.

LEMMA 1. *Let $\mathcal{B}$ be any set of complete bipartite graphs and let $\mathcal{B}'$ be the kernel of $\mathcal{B}$. Then for any graph H, H admits a $\mathcal{B}$-factor if and only if H admits a $\mathcal{B}'$-factor.* □

The main result of this section can now be stated precisely.

THEOREM 4. *Let $\mathcal{B}$ be an irreducible set of complete bipartite graphs. If $\mathcal{B}$ is not a sequential set of stars, then the $\mathcal{B}$-factor problem is $\mathcal{NP}$-hard.*

The remainder of this section outlines a proof of Theorem 4. Some details are left to the reader. For the remainder of this section let $\mathcal{B}$ be an arbitrary irreducible set of complete bipartite graphs and let $\mathcal{B}_s=\mathcal{B}\cap\{\mathcal{S}_i\colon i\geqq 1\}$, the set of stars in $\mathcal{B}$.

We first observe that the proof of Theorem 3 generalizes directly to the following:

THEOREM 3′. *If $\mathcal{B}_s$ is a nonsequential set of stars, then the $\mathcal{B}$-factor problem is $\mathcal{NP}$-hard.*

It remains to consider the case where $\mathcal{B}_s$ is a sequential set of stars and $\mathcal{B}$ contains at least one nonstar. Our $\mathcal{NP}$-hardness proof constructs a bipartite graph whose only $\mathcal{B}$-factors are in fact $K_{s,t}$-factors for some specified $K_{s,t}\in\mathcal{B}$. Of central importance in establishing this as well as other properties is the following simple observation.

LEMMA 2. *If H is any bipartite graph with parts of sizes p,q where $p\leqq q$, and if H admits a $K_{s,t}$-factor, where $s\leqq t$, then $q/p\leqq t/s$.*

Our $\mathcal{NP}$-hardness proof makes use of the following notation first presented in [28], [29].

A *module* is a graph M together with a nonempty subset $C\subseteq V(M)$ of distinguished vertices that we call the *connector* vertices of M. (The elements of $V(M)-C$ are called the *interior* vertices of M). If G is any fixed graph, then M is said to be a *G-module* if M admits a G-packing saturating all of its interior vertices (plus some, possibly empty, subset of its connector vertices).

A graph H is a *modular extension* of the module M if H contains M as an induced subgraph in which no interior vertex of M is adjacent to a vertex of $H-M$ (that is, M is connected to the rest of H only through its connector vertices). Let $\pi=\{G_1,\cdots,G_d\}$ be any G-packing of some modular extension H of M. A vertex of v of M is said to be *bound* to M by π, if $v\in V(G_i)$ implies $V(G_i)\subseteq V(M)$. A G-module M is *internally G-coherent* if every G-factor of every modular extension of M binds to M all of its interior vertices.

Now suppose $\mathcal{B}_s$ is a sequential set $\{S_1,\cdots,S_q\}$ of stars and that $\mathcal{B}$ contains at least one nonstar. We say that a nonstar $K_{s,t}$ with $1<s\leqq t$, is *minimal* in $\mathcal{B}$ if $K_{s,t}\in\mathcal{B}$ and for all $K_{s',t'}\in\mathcal{B}$ with $s'\leqq t'$ either $s'=1$, or $s'>s$, or $t'>t$. By the irreducibility of $\mathcal{B}$, it follows that $t>qs$.

Suppose $K_{s,t}$ is minimal in $\mathcal{B}$. An *s,t-fork* is a complete bipartite graph $K_{s+1,t}$ in which three vertices in the part of size $s+1$ are distinguished as connector vertices. Similarly an *s,t-cross* is a complete bipartite graph $K_{s+2,t}$ in which four vertices in the part of size $s+2$ are distinguished as connector vertices.

We find it convenient to depict an s,t-fork and s,t-cross schematically as in Fig. 5.

It should be clear that both s, t-forks and s, t-crosses are internally $K_{s,t}$-coherent, and that an $K_{s,t}$-factor of any modular extension of either one binds to that module

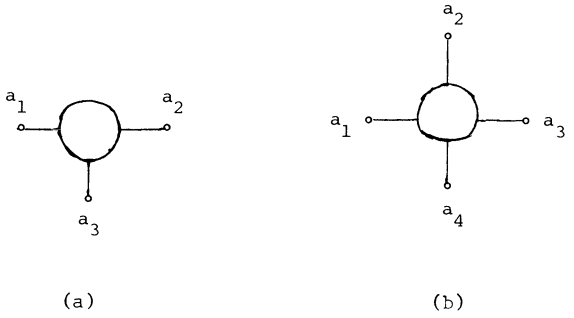


FIG. 5. *An s, t-fork* (a) *and s, t-cross* (b).

its interior vertices plus exactly two of its connector vertices. If we say that in such a situation the unbound connector vertices are *chosen*, then a fork (respectively, a cross) can be used to force an arbitrary choice of one out of three (respectively, two out of four) vertices. We use s, t-forks and crosses to build larger modules with similar properties.

We define a *c by d s, t-array* to be the graph formed by connecting s, t-forks and s, t-crosses into a c by d array as shown in Fig. 6.

A c by d s, t-array can be viewed as a $K_{s,t}$-module with connector vertices $x_1, \cdots, x_d$. The following lemma summarizes the relevant properties of such modules.

LEMMA 3. *Let M be a c by d s, t-array with $c \leqq d$. Then*

(a) *Any $K_{s,t}$-factor of any modular extension of M binds to M its internal vertices plus at least $d-c$ of its connector vertices*: *and*

(b) *The graph M minus any c of its connector vertices admits a $K_{s,t}$-factor.*

Proof. (a) It is easy to confirm that M is a bipartite graph with sizes $cds+c$ and cdt, where all of the connectors belong to the part of size $cds+c$. The result follows by Lemma 2.

(b) Suppose $x_{i_1}, x_{i_2}, \cdots, x_{i_c}$ are the chosen connector vertices, where $i_1 < i_2 < \cdots < i_c$. Then the module in array position (j, k) binds

(i) its left and right connectors if $k = i_j$;

(ii) its left and lower connectors if $k < i_j$;

(iii) its right and upper connectors if $k = i_t$, $t > j$; and

(iv) its right and lower connectors if $k > i_j$ and $k \neq i_t$, $t > j$.

The resulting $K_{s,t}$-packing saturates all of the vertices of M except the connectors $x_{i_1}, \cdots, x_{i_c}$. □

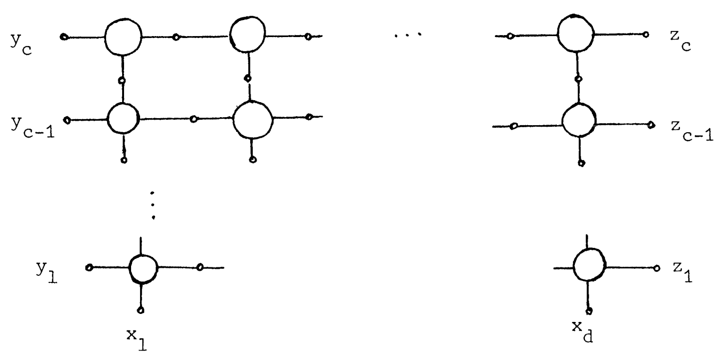


FIG. 6. *A c by d s, t-array.*

It follows from Lemma 3 that, with regard to $K_{s,t}$-factors, c by d s, t-arrays, wherever they appear as modules in a larger graph, force a choice of at most c of their d connector vertices. This choice is exploited in the following reduction of the t-dimensional matching problem to the $\mathscr{B}$-factor problem.

Let n be an integer and let $P\subseteq\{1,2,\cdots,n\}^t$. Suppose $P=\{\langle p_1^i, p_2^i,\cdots,p_t^i\rangle\colon 1\leqq i\leqq l\}$ and consider the graph $H(s,t,P)$, depicted in Fig. 7, in which each x_j^i, $1\leqq i\leqq l$, $1\leqq j\leqq s$, is adjacent to each of (p_r^i, r), $1\leqq r\leqq t$.

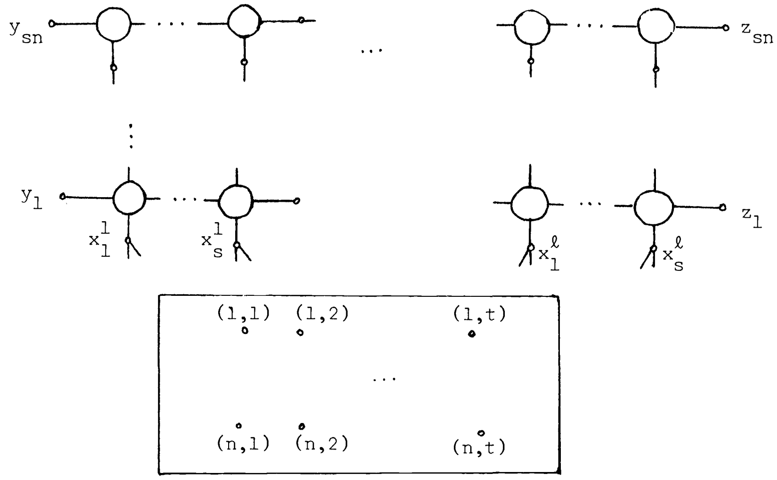


FIG. 7. *The graph* $H(s,t,P)$.

$H(s,t,P)$ is clearly a modular extension of an sn by sl s, t-array. Our concentration on $K_{s,t}$-factors is justified by the following lemma.

LEMMA 4. *$H(s,t,P)$ admits a $\mathscr{B}$-factor if and only if $H(s,t,P)$ admits a $K_{s,t}$-factor.*

Proof. $H(s,t,P)$ is easily seen to be a bipartite graph with parts of size s^3nl+sn and $s^2tnl+tn$. Furthermore, no set of s vertices in the part of size s^3nl+sn has more than t common neighbors. This, together with the minimality of $K_{s,t}$ in $\mathscr{B}$ implies that no bipartite graph with a ratio of part sizes at least t/s except $K_{s,t}$ appears in any $\mathscr{B}$-factor of $H(s,t,P)$. But since the average ratio of the part sizes of the bipartite graphs in any $\mathscr{B}$-factor of $H(s,t,P)$ is t/s, it follows that any $\mathscr{B}$-factor must be a $K_{s,t}$-factor. □

As a consequence of Lemma 4 the proof of Theorem 4 is completed with the following lemma.

LEMMA 5. *$H(s,t,P)$ admits a $K_{s,t}$-factor if and only if P admits a t-dimensional matching.*

Proof. Suppose $\{\langle p_1^j,\cdots,p_t^j\rangle\colon j\in\{i_1,\cdots,i_n\}\}$ is a t-dimensional matching of P. Let F be any $K_{s,t}$-factor of the array-module of $H(s,t,P)$ minus its connectors x_j^i, $i\in\{i_1,\cdots,i_n\}$, $1\leqq j\leqq s$ (the existence of which is guaranteed by Lemma 3). Then F together with the n subgraphs induced on the vertex sets $\{x_1^i,\cdots,x_s^i,(p_1^i,1),(p_2^i,2),\cdots,(p_t^i,t)\}$ for $i\in\{i_1,\cdots,i_n\}$, forms a $K_{s,t}$-factor of $H(s,t,P)$.

Conversely, suppose $H(s,t,P)$ admits a $K_{s,t}$-factor F. By Lemma 3, F binds all but at most sn of the connector vertices to the array-module of $H(s,t,P)$. But $H(s,t,P)$ minus the internal vertices of its array module is easily seen to be a bipartite graph of sizes ln and tn, and hence F restricted to this graph must induce a partition of

$\{(j, i): 1 \leqq i \leqq t, 1 \leqq j \leqq n\}$ into exactly n t-tuples each belonging to P—i.e., a t-dimensional matching of P. □

Acknowledgment. We wish to express our sincere thanks to an anonymous referee for his/her critical review of an earlier version of this paper.

REFERENCES

[1] J. AKIYAMA AND M. KANO, *Path factors of a graph*, preprint, 1983.

[2] J. AKIYAMA, D. AVIS AND H. ERA, *On a* $\{1, 2\}$*-factor of a graph*, TRU Math., 16 (1980), pp. 97-102.

[3] A. AMAHASHI AND M. KANO, *On factors with given components*, Discrete Math., 42 (1982), pp. 7-26.

[4] H. B. BELCK, *Reguläre Factoren von Graphen*, J. Reine Angew. Math., 188 (1950), pp. 228-252.

[5] C. BERGE, *Two theorems in graph theory*, Proc. Nat. Acad. Sciences, USA, 43 (1957), pp. 842-844.

[6] ——— *Sur le couplage maximum d'un graphe*, C.R. Acad. Sciences Paris, 247 (1958), pp. 258-259.

[7] C. BERGE AND M. LAS VERGNAS, *On the existence of subgraphs with degree constraints*, Proc. Nederl. Akad. Wetensch, (cited in [31]).

[8] G. CORNUEJOLS AND W. PULLEYBLANK, *A matching problem with side conditions*, Discrete Math., 29 (1980), pp. 135-159.

[9] G. CORNUEJOLS, D. HARTVIGSEN AND W. PULLEYBLANK, *Packing subgraphs in a graph*, OR Letters, 4 (1982), pp. 139-143.

[10] G. CORNUEJOLS AND W. R. PULLEYBLANK, *Critical graphs, matchings and tours*, Combinatorica (1983), pp. 35-51.

[11] E. A. DINIC, *Algorithm for solution of a problem of maximum flow in a network with power estimation*, Soviet Math. Dokl., 11 (1970), pp. 1277-1280.

[12] J. EDMONDS, *Paths, trees, and flowers*, Canad. J. Math., 17 (1965), pp. 449-467.

[13] ——— *Maximum matching and a polyhedron with* $(0, 1)$ *vertices*, J. Res. Nat. Bureau of Standards, 69B (1965), pp. 125-130.

[14] J. EDMONDS AND E. L. JOHNSON, *Matching: a well-solved class of integer linear programs*, in Combinatorial Structures and Their Applications, R. K. Guy et al., eds., Gordon and Breach, N.Y., 1970, pp. 89-92.

[15] P. ERDÖS AND T. GALLAI, *On the minimum number of vertices representing the edges of a graph*, Publ. Math. Inst. Hung. Acad. Sci., 6 (1961), pp. 181-203.

[16] H. N. GABOW, *An efficient reduction technique for degree constrained subgraph and bidirected network flow problems*, Proc. 15th Annual ACM Symposium on Theory of Computing, pp. 448-456.

[17] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability*, W. H. Freeman, San Francisco, 1979.

[18] A. HAJNAL, *A theorem on k-saturated graphs*, Canad. J. Math., 17 (1965), pp. 720-772.

[19] P. HELL, Summer Research Workshop in Combinatorics, Simon Fraser Univ., Aug. 8, 1979; Applied Discrete Math. Seminar, Rutgers Univ., Oct. 23, 1979; 12th Southeastern Conf. on Combinatorics, Graph Theory, and Computing, L.S.U., March 4, 1981; Math. Programming and Combinatorics Seminar, Princeton Univ., Nov. 11, 1981.

[20] P. HELL AND D. G. KIRKPATRICK, *Scheduling, matching and coloring*, Colloq. Math. Soc. János Bolyai, 25, Algebraic Methods in Graph Theory, Szëged, Hungary, North-Holland, Amsterdam, 1978, pp. 272-280.

[21] ———, *On generalized matching problems*, Inform. Proc. Lett., 12 (1981), pp. 33-35.

[22] ———, *Star factors and star packings*, TR 82-6, Dept. Computing Science, Simon Fraser University, Burnaby, British Columbia 1982.

[23] ———, *Packings by cliques and by finite families of graphs*, Discrete Math., 49 (1984), pp. 118-133.

[24] ———, *(g, f)-factors*, to appear.

[25] J. E. HOPCROFT AND J. D. ULLMAN, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, Reading, MA, 1979.

[26] J. E. HOPCROFT AND R. M. KARP, *An* $n^{5/2}$ *algorithm for maximum matchings in bipartite graphs*, SIAM J. Comput., 2 (1973), pp. 225-231.

[27] T. KAMEDA, *Testing deadlock-freedom of computer systems*, J. Assoc. Comput. Mech., 27 (1980), pp. 270-280.

[28] D. G. KIRKPATRICK AND P. HELL, *On the completeness of a generalized matching problem*, in Proc. Tenth Annual ACM Symposium on Theory of Computing, 1978, pp. 240-245.

[29] ———, *On the complexity of general graph factor problems*, SIAM J. Comput., 12 (1983), pp. 601-609.

[30] D. KÖNIG, *Graphen and Matrizen*, Mat. Fiz. Lapok, 38 (1931), pp. 116-119.

[31] M. LAS VERGNAS, *An extension of Tutte's 1-factor theorem*, Discrete Math., 23 (1978), pp. 241-255.

[32] L. LOVÁSZ, *Subgraphs with prescribed valencies*, J. Combin. Theory, 19B (1975), pp. 269-271.

[33] S. MICALI AND V. V. VAZIRANI, *An $O(\sqrt{|V|}\cdot|E|)$ algorithm for finding maximum matching in general graphs*, Proc. 21st Annual Symposium on Foundation of Computer Sciences 1980, pp. 17-27.

[34] J. MÜHLBACHER, *F-factors of graphs: a generalized matching problem*, Inform Proc. Lett., 8 (1979), pp. 207-214.

[35] Y. SHILOACH, *Another look at the degree constrained subgraph problem*, Inform. Proc. Lett., 12 (1981), pp. 9-92.

[36] W. T. TUTTE, *The factorisation of linear graphs*, J. London Math. Soc., 22 (1947), pp. 107-111.

[37] ———, *The factors of graphs*, Canad. J. Math., 4 (1952), pp. 314-328.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

006

VALUES OF GRAPH-RESTRICTED GAMES*

GUILLERMO OWEN†

Abstract. We consider the problem of modifying n-person games so as to take account of the difficulties imposed by lack of communications, and the opportunities this might accord to intermediaries.

In this model, the members of a finite set are simultaneously players in a game and vertices of a graph. A combination of these two structures gives rise to a new, modified game in which the only effective coalitions are those corresponding to connected partial graphs. We study the relationship between the power indices of the original game and the restricted game; for the special case where the graph is a tree, this relationship is especially easy to analyze.

Several examples are studied in detail.

Key words. game theory, graphs, Shapley value, centrality

AMS(MOS) subject classification. 90

Preliminary. Let $N=\{1, 2, \cdots, n\}$ be a finite set. In this paper the elements of N will be, simultaneously, the players in a game v and the nodes in a graph Γ.

The *game* v is merely a *characteristic function.* Thus v is a mapping from 2^N, the set of subsets of N, into the reals, satisfying

$$v(\emptyset)=0. \tag{1}$$

If, moreover, v satisfies the condition

$$v(S\cup T)\geqq v(S)+v(T) \quad \text{if } S\cap T=\emptyset, \tag{2}$$

we shall say v is a *proper* or *super-additive* game. The space G_N of all n-person games is a (2^n-1)-dimensional real vector space; the set of all proper games is a full-dimensional cone Q_N in G_N.

The *graph* Γ is a pair (N, A), where N is as above and A is a collection of pairs $\{i, j\}$, $i\neq j$, $i, j\in N$. The pairs are unordered, so that the graph is *undirected.* The members of A are *arcs* of the graph Γ.

If S is a subset of N, we define A_S to be the set of all pairs $\{i, j\}$ where $\{i, j\}\in A$ and $i, j\in S$. The *partial graph* Γ_S is then the pair (S, A_S). Thus the partial graph includes all nodes in S and all arcs which lie between such nodes. We shall say S is *connected* (in Γ) if the partial graph Γ_S is connected. The *components of* S are the maximal connected subsets of S (in Γ).

1. The restricted game.

In what follows, we shall assume Γ is a fixed but otherwise arbitrary graph. The game v is also arbitrary (but not fixed); no relationship is assumed between v and Γ other than that the players in the former are nodes in the latter.

Heuristically, we shall assume that the game v represents the economic capabilities of the players, i.e., the coalition S can obtain utility $v(S)$ only if its members can come to an agreement to cooperate. Unfortunately, this may not be possible since there may be no lines of communications open among them. The graph Γ, now, represents the communications channels available: i can communicate directly with j if and only if $\{i, j\}\in A$. Of course, even if $\{i, j\}\notin A$, it may still be possible for i to communicate with

* Received by the editors August 14, 1984, and in revised form February 1, 1985.

† Department of Mathematics, Naval Postgraduate School, Monterey, California 93943 and School of Social Sciences, University of California, Irvine, California 92717.

j. This will, however, require the cooperation of some intermediaries who can relay a message, i.e., players who define a path, in Γ, from i to j.

With this in mind, we can now define the graph-restricted game, v/Γ, as a game w (with the same players) given by

$$w(S) = \begin{cases} v(S) & \text{if } S \text{ is connected,} \\ \sum_{k=1}^{m} v(T_k) & \text{otherwise,} \end{cases} \tag{3}$$

where $T_1, T_2, \cdots, T_m$ are the components of S (in Γ).

It is easy to see that the mapping L_Γ defined by $L_\Gamma(v) = v/\Gamma$ is a linear mapping of G_N into itself. It is not so obvious that L_Γ preserves the cone Q_N of super-additive games. We prove this in the Appendix.

THEOREM 1. *The mapping L_Γ is a linear mapping of G_N into itself, preserving the cone Q_N.*

Clearly, L_Γ is not full-dimensional (its kernel is not zero). It is of interest to find its image. To do this, we first define a basis, the *unanimity game basis*, for G_N.

If T is an arbitrary (nonempty) subset of N, the game u_T, defined by

$$u_T(S) = \begin{cases} 1 & \text{if } T \subset S, \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

is the *unanimity game on T.* There are 2^{n-1} unanimity games in G_N, and they can be shown to form a basis (see e.g. Owen [1982]).

Since the unanimity games form a basis, each game v can be expressed as a linear combination of them:

$$v = \sum_T \Delta_v(T) u_T.$$

The coefficients in this linear combination are given by the formula

$$\Delta_v(S) = \sum_{T \subset S} (-1)^{s-t} v(T) \tag{5}$$

where s, t are the cardinalities of S and T. Following Harsanyi [1958], we shall call them the *dividends* in game v. They satisfy, of course,

$$v(S) = \sum_{T \subset S} \Delta_v(T). \tag{6}$$

We note that the two systems (5) and (6) are equivalent, i.e. any set of numbers $\Delta_v(T)$ which satisfy (6) necessarily satisfy (5) as well.

To study the behavior of the mapping L_Γ, we will consider the images of the basic games u_T. In fact, if T is connected (in Γ), it is not too difficult to see that $L_\Gamma(u_T) = u_T$. If T is not connected, however, we find that $L_\Gamma(u_T) \equiv u_T/\Gamma$ is given by c_T, where

$$c_T(S) = \begin{cases} 1 & \text{if there is a connected set } K \text{ such that } T \subset K \subset S, \\ 0 & \text{otherwise.} \end{cases}$$

Just as u_T is usually called the unanimity game on T, so c_T could be called the "connect T in Γ" game. It is now necessary to express c_T in terms of the unanimity games.

$$c_T = \sum_S \Delta_{c_T}(S) u_S. \tag{7}$$

Calculation of the dividends $\Delta_{c_T}(S)$ is not trivial. It is easy to see, however, that $\Delta_{c_T}(S) = 0$ unless $T \subset S$. Not so easy, but very important, is the next result.

THEOREM 2. *If S is disconnected, then $\Delta_{c_T}(S)=0$ for all T.*

The proof is in the Appendix.

An important consequence of Theorem 2 is

THEOREM 3. *The image of the mapping L_Γ is the set spanned by the unanimity games u_S, where S is connected in Γ. These games form a basis for the image,* $\mathrm{IM}(L_\Gamma)$.

Proof. By Theorem 2, the image $L_\Gamma(u_T)$ of any basic game is a linear combination of games u_S, where S is connected. Moreover, u_S is its own image whenever S is connected, so each u_S belongs to the image space. It follows that these u_S span the image space. Since they are known to be independent, they form a basis.

A further important consequence is a formula which relates the dividents Δ_v in the original unrestricted game to those, Δ_w, in the restricted game. We have, in fact,

$$\Delta_w(S)=\sum_{T\subset S}\Delta_{c_T}(S)\Delta_v(T). \tag{8}$$

The validity of (8) follows from the fact that the Δ_v are the coefficients of v in terms of the u_T, while the Δ_{c_T} are the coefficients of the images of the u_T in terms of the u_S. The result follows from the linearity of the mapping.

2. The power indices. In [1977] Myerson, treating these games, suggested that their Shapley value should be studied. We will not discuss his arguments here, but propose to look at the Shapley value (Shapley [1953]) and, to a lesser extent, at the Banzhaf-Coleman value (Banzhaf, [1965], Coleman [1971]) of the games v/Γ as well as its relation to the value (or index) for the original game v.

As is well known, the Shapley value of a game v is given by the formula

$$\phi_i[v]=\sum_{\substack{S\subset N\\ i\in S}}\frac{s!(n-s-1)!}{n!}[v(S\cup\{i\})-v(S)]. \tag{9}$$

In terms of the dividends, however, the value has the simpler form

$$\phi_i[v]=\sum_{\substack{S\subset N\\ i\in S}}\frac{1}{s}\Delta_v(S). \tag{10}$$

Similarly, the Banzhaf-Coleman index can be expressed as

$$\psi_i[v]=\sum_{\substack{S\subset N\\ i\in S}}2^{1-s}\Delta_v(S). \tag{11}$$

If we think of either ϕ or ψ as a linear mapping from G_N into R^N, expressions (10) and (11) tell us that the mapping is most easily handled in terms of the unanimity base.

Unfortunately, of course, computation of the dividends $\Delta_v(S)$, for an arbitrary game v, is usually quite lengthy. An attempt, then, to compute either the Shapley value or the Banzhaf-Coleman index for an arbitrary v/Γ through the use of equations (8), (10), and (11) would probably require too many computations. There are situations, however, when both the unrestricted game, v, and the communications network, Γ, are easy enough to analyze. In such cases, use of (8), (10), and (11) is probably the fastest way to compute the indices of power.

3. Trees. Of particular interest is the case where Γ is a tree, i.e., a connected graph with no circuits. (A circuit is defined in the usual manner as a sequence $i_1, i_2, i_3, \cdots, i_k$ of $k\geqq 3$ nodes such that $\{i_1, i_2\}, \{i_2, i_3\}, \cdots, \{i_{k-1}, i_k\}, \{i_k, i_1\}$ all belong to A.) In this case, computation of the power indices is simplified by the following results.

THEOREM 4. *In a tree, the intersection of connected subgraphs is connected.*

The proof appears in the Appendix. An important corollary is

THEOREM 5. *Let Γ be a tree. For any $T \subset N$, there exists a unique minimal* (*in the sense of set inclusion*) *connected S such that $T \subset S$.*

Proof. Let us define

$$H(T) = \bigcap \{S \mid T \subset S,\ S \text{ is connected}\}.$$

Clearly $T \subset H(T)$ and, by Theorem 4, $H(T)$ is connected. Also, by definition, $H(T)$ is a subset of any connect S with $T \subset S$, and so is minimal.

We shall call $H(T)$ the (*connected*) *hull* of T. As an example, in the tree of Fig. 1, the hull of $\{1, 7, 8, 12\}$ is $\{1, 2, 6, 7, 8, 11, 12\}$. Note, moreover, that $S = H(S)$ if and only if S is connected.

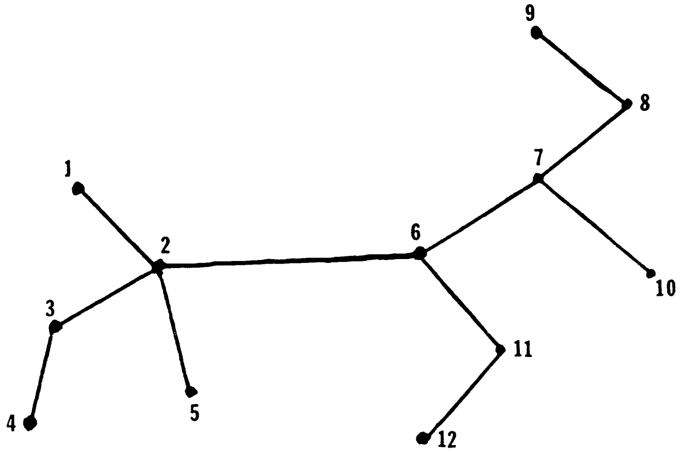


FIG. 1

The importance of this lies in

THEOREM 6. *If Γ is a tree, then $L_\Gamma(u_T) = u_{H(T)}$.*

Proof. We have already seen that $L_\Gamma(u_T) = c_T$. That $u_{H(T)} = c_T$ follows directly from their definitions.

The importance of Theorem 6 is due to the fact that the dividends Δ_{c_T} take the simple form

$$\Delta_{c_T}(S) = \begin{cases} 1 & \text{if } S = H(T), \\ 0 & \textit{otherwise}, \end{cases}$$

so that (still assuming Γ is a tree) the formula (8) takes the form

$$\Delta_w(S) = \sum_{\substack{T \\ H(T)=S}} \Delta_v(T). \tag{12}$$

This can be further simplified if we ask for the conditions that make $H(T) = S$. We will say that $i \in S$ is an *extreme point* of S if, in the partial graph Γ_S, i has order 0 or 1 (i.e. there is at most one other $j \in S$ which is directly connected to i). Let $E(S)$ denote the set of extreme points of S. Now we have

THEOREM 7. *If Γ is a tree, then a necessary and sufficient condition for $H(T) = S$ is that*

(i) *S is connected,*

(ii) $T \subset S$,

(iii) $E(S) \subset T$.

The proof of Theorem 7 appears in the Appendix. (12) further reduces to

$$\Delta_w(S) = \sum_{\substack{T \\ E(S) \subset T \subset S}} \Delta_v(T). \tag{13}$$

This is a considerably simpler expression. In Fig. 1, for example, the set $S = \{1, 2, 6, 7, 8, 11, 12\}$ is connected. Now $E(S) = \{1, 8, 12\}$ and so $S = H(T)$ if and only if T contains 1, 8, 12 and some subset of $\{2, 6, 7, 11\}$. It is clear that there are $2^4 = 16$ such sets T.

4. Examples. We consider several cases here. In each case, Γ is assumed to be a tree.

(a) *Inessential games.* Let $\mu = (\mu_1, \cdots, \mu_n)$ be a vector. The game v, given by

$$v(S) = \sum_{i \in S} \mu_i$$

is an inessential game.

It is not too difficult to see that, in this case, $v/\Gamma = v$. In fact, we have here $\Delta_v(\{i\}) = \mu_i$ for all i; $\Delta_v(S) = 0$ for all other S. Since the sets $\{i\}$ are trivially connected, v and w will coincide. The Shapley value is of course given by $\phi_i[v] = \mu_i$ for all $i \in N$. This is also the Banzhaf-Coleman index. (All this will still be true even if Γ is not a tree.)

(b) *Unanimity games.* The unanimity games have already been defined. We have also seen that, whenever Γ is a tree, $u_T/\Gamma = u_{H(T)}$. The values here are easily obtained:

$$\phi_i[u_T] = \begin{cases} 1/t & \text{if } i \in T, \\ 0 & \text{if } i \notin T, \end{cases}$$

$$\phi_i[u_T/\Gamma] = \begin{cases} 1/h & \text{if } i \in H(T), \\ 0 & \text{if } i \notin H(T), \end{cases}$$

where t, h are the cardinalities of T, $H(T)$ respectively.

In case Γ is not a tree, the results can be very complicated.

(c) "*Pure overhead*" *games.* Let T be an arbitrary (nonempty) subset of N. The pure overhead game on T, p_T, is defined by

$$p_T(S) = \begin{cases} -1 & \text{if } S \cap T \neq \emptyset, \\ 0 & \text{if } S \cap T = \emptyset. \end{cases}$$

Thus, p_T is the negative dual of the unanimity game u_T; its Shapley value is

$$\phi_i[p_T] = \begin{cases} -1/t & \text{if } i \in T, \\ 0 & \text{if } i \notin T. \end{cases}$$

To analyze the game p_T/Γ, we first note that, if S is not a subset of T, or if $S = \emptyset$, $\Delta_{p_T}(S) = 0$. Assuming S is a nonempty subset of T, we have

$$\Delta_{p_T}(S) = \sum_{K \subset S} (-1)^{s-k} p_T(K)$$

and, since $p_T(K) = -1$ for all $K \subset S$, $K \neq \emptyset$, this is

$$\Delta_{p_T}(S) = -\sum_{\substack{K \subset S \\ K \neq \emptyset}} (-1)^{s-k} = -\sum_{k=1}^{s} (-1)^{s-k} \binom{s}{k}$$

or $(-1)^s$. (See the Appendix for this.) Thus

$$\Delta_{p_T}(S)=\begin{cases}(-1)^s & \text{if } S\subset T, S\neq\emptyset,\\ 0 & \text{otherwise.}\end{cases}$$

The graph-restricted game, p_T/Γ, is given by $w(S)=m$, where m is the number of components of S which have a nonempty intersection with T.

Application of (13) gives us, for connected sets S,

$$\Delta_w(S)=\sum_{\substack{K\\ E(S)\subset K\subset S\cap T}}\Delta_{p_T}(K).$$

Since $\Delta_{p_T}(K)=0$ unless $K\subset T$, we see that $\Delta_w(S)=0$ unless $E(S)\subset T$. Assuming that this holds (and also that S is connected and nonempty), we will have

$$\Delta w(S)=\sum_{\substack{K\\ E(S)\subset K\subset S\cap T}}(-1)^k.$$

Suppose $E(S)$ has e elements and $S\cap T$ has s' elements. Then

$$\Delta_w(S)=\sum_{k=e}^{s'}(-1)^k\binom{s'-e}{k-e}.$$

This expression equals $(-1)^e$ if $s'=e$ and 0 otherwise. Thus we conclude that $\Delta_w(S)=0$ unless $E(S)=S\cap T$, i.e. *all the extreme points of S, and none of its interior points belong to T.* For such S, $\Delta_w(S)=(-1)^e$, except that $\Delta_w(\emptyset)=0$.

A special, very simple case occurs when $T=N$. In this case, $\Delta_w(S)$ will be zero unless $S=E(S)$, i.e. all points of S are extreme. For connected S, this can only happen if S reduces to either a single point $\{i\}$ or an arc $\{i,j\}$. We would have, in this case,

$$\Delta_w(S)=\begin{cases}-1 & \text{if } S=\{i\},\\ 1 & \text{if } S=\{i,j\}\in A,\\ 0 & \text{otherwise.}\end{cases}$$

Application of formula (10) gives us

$$\phi_i[w]=-1+\frac{d_i}{2}$$

where d_i is the degree of node i in the graph Γ, i.e. the number of arcs incident on i.

In the more general case, where T is a proper subset of N, computation is more complicated and requires some combinatorial arguments. As an example, consider the network in Fig. 2, where $n=13$, and $T=\{1, 2, 3, 4, 6, 7, 10, 11, 13\}$.

A game such as this is best handled by subdividing the graph into the partial graphs corresponding to $S_1=\{1,2,3,4,5\}$, $S_2=\{2,6,7,8\}$ and $S_3=\{6,9,10,11,12,13\}$, i.e. the graph is split at its two interior "costly" nodes, 2 and 6. (The point is that any S with $\Delta_w(S)\neq 0$ must be a subset of one of these, as it must be connected and have no interior costly nodes.) Once this is done, we can compute ϕ, with some work. For $i=1$, we note that the only S such that $i\in S$ and $\Delta_w(S)\neq 0$ are (a) $\{i\}$ alone, and (b) sets which include 1, 5, and at least one of 2, 3, and 4. Of case (b), there are three which include one, three which include two, and one which includes all three of these nodes. We have then

$$\phi_1[w]=\tfrac{1}{1}(-1)+\tfrac{1}{3}(3)(-1)^2+\tfrac{1}{4}(3)(-1)^3+\tfrac{1}{5}(-1)^4=-\tfrac{11}{20}.$$

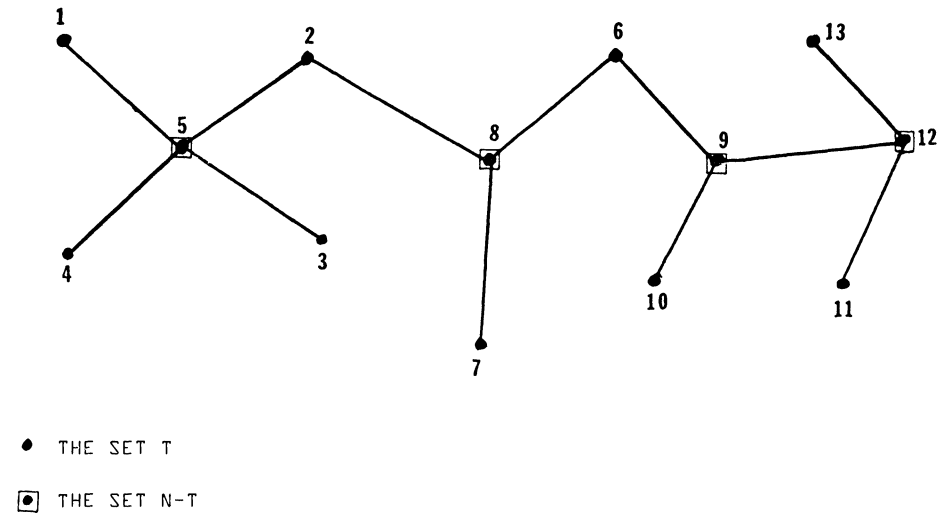


FIG. 2

Similar calculations will give us the value vector

$$\phi[w]=(-\tfrac{11}{20}, -\tfrac{2}{5}, -\tfrac{11}{20}, -\tfrac{11}{20}, \tfrac{6}{5}, -\tfrac{11}{60}, -\tfrac{7}{12}, \tfrac{3}{4}, \tfrac{7}{10}, -\tfrac{3}{5}, -\tfrac{3}{5}, \tfrac{7}{10}, -\tfrac{3}{5}).$$

(d) *Quadratic measure games.* If $\mu=(\mu_1, \mu_2, \cdots, \mu_n)$ is a nonnegative vector, the set function

$$\mu(S)=\sum_{i\in S} \mu_i$$

is of course a measure. Any game which can be expressed as a quadratic function of such a measure (or even of several such measures) is a quadratic measure game.

Let us write

$$v(S)=\left(\sum_{i\in S} \mu_i\right)^2-\sum_{i\in S} \mu_i^2.$$

This is as general a function as we need consider since any quadratic can be written as a linear combination of such functions, plus or minus an additive set function. The specific form chosen has the advantage of being 0-normalized, i.e. $v(\{i\})=0$ for any i.

Games such as this have an interesting property, namely that the only coalitions with nonzero dividends are the two-person sets. In fact, it is easily checked that

$$\Delta_v(S)=\begin{cases} 2\mu_i\mu_j & \text{if } S=\{i,j\}, \quad i\neq j, \\ 0 & \text{otherwise.} \end{cases}$$

From this the Shapley value is easily computed; we have

$$\phi_i[v]=\sum_{j\neq i} \tfrac{1}{2}(2\mu_i\mu_j)$$

and so

$$\phi_i[v]=\mu_i\mu(N)-\mu_i^2.$$

The graph-restricted game is best analyzed by using (13). Since $\Delta_v(T)=0$ unless T has exactly two elements, it will follow that $\Delta_w(S)=0$ unless S has at most two extreme points. But if S has only one extreme point, it has only one point and so

$\Delta_w(S)=0$. Thus $\Delta_w(S)$ is zero except for connected sets with two extreme points. In a tree, such an S is necessarily a path. We have

$$\Delta_w(S)=\begin{cases}2\mu_j\mu_k & \text{if } S \text{ is the path from } j \text{ to } k,\\ 0 & \text{if } S \text{ is not a path.}\end{cases}$$

Application of (10) now gives us

$$\phi_i[w]=2\sum\frac{\mu_j\mu_k}{d(j,k)+1}$$

where the sum is taken over all pairs j, k such that i lies on the path from j to k, and $d(j,k)$ is the distance from j to k (so that the denominator is in fact equal to the number of nodes on the path).

The computation of this last sum can now be readily carried out, even for large trees, by purely combinatorial methods, i.e. generating functions. As an example, consider the tree of Fig. 3, letting the measure vector give weight 1 to each node. With this measure, v takes the simple form $v(S)=s(s-1)$ where s is the cardinality of S. Thus $v(N)=156$.

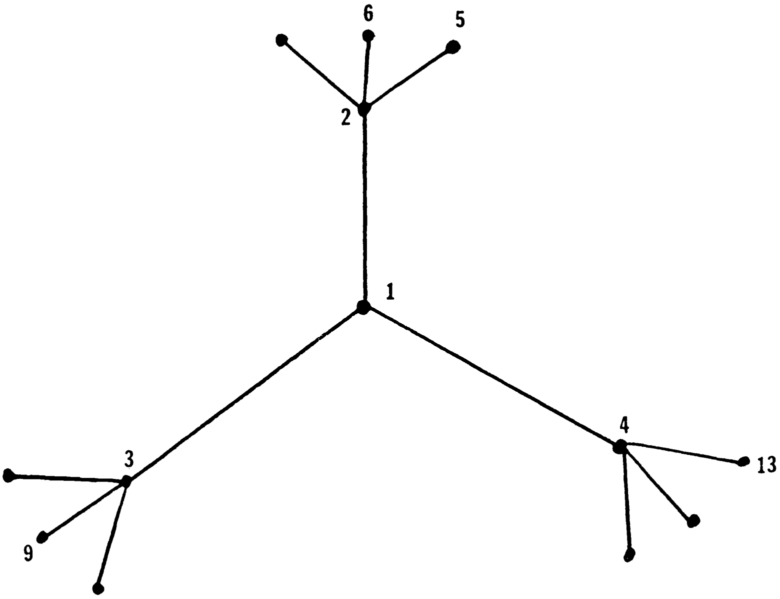


FIG. 3

To compute $\phi[w]$, we note that, if we were to cut the three arcs incident on node 1, the graph would have four components: $S_0=\{1\}$, $S_1=\{2,5,6,7\}$, $S_2=\{3,8,9,10\}$, $S_3=\{4,11,12,13\}$. For each component, we form the generating function which gives the number of nodes lying at a given distance from node 1. Thus S_0 has one node at distance 0; the others have each one node at distance 1, and three at distance 2, and so

$$\theta_0=1, \qquad \theta_1=\theta_2=\theta_3=x+3x^2.$$

Next, we compute

$$F_1(x)=\left(\sum\theta_r\right)^2-\sum(\theta_r)^2=(1+3x+9x^2)^2-1-3(x+3x^2)^2. \tag{14}$$

Thus

$$F_1(x)=6x+24x^2+36x^3+54x^4.$$

This is the generating function for the number of paths of a given length passing through node 1. (Each path is actually counted twice, once in each direction.) To see why formula (14) works, we note that any path through node 1 must start in one of the four components S_0, S_1, S_2, S_3 and terminate in another; its length will be the sum of the distances of its two end-points from 1. The "square-of-the-sum" term in (14) considers all paths to and from 1 and adds their lengths. This counts too much, however, as it will include pairs of paths to 1 from nodes in the same component. Hence, the "sum-of-the-squares" term is subtracted so as to remove all these unwanted pairs.

It is now trivial to evaluate ϕ_1; by (12), it will be

$$\phi_1[w]=\tfrac{6}{2}+\tfrac{24}{3}+\tfrac{36}{4}+\tfrac{54}{5}=30.8.$$

Alternatively, game theorists may notice that $F_1(x)$ is nothing other than the partial derivative, with respect to x_1, of the multilinear extension (see Owen [1972]) of the game w, evaluated at the point $(x, x, \cdots, x)$. The value can be obtained by integrating this from 0 to 1, giving, of course, the same result.

Similar considerations will give us

$$\phi_2=\phi_3=\phi_4=22.5633, \qquad \phi_5=\phi_6=\cdots=\phi_{13}=6.4.$$

In case a different measure were used, the procedure would be the same, except that the generating functions $\theta_r(x)$ would give, not the number of nodes at a given distance from the node being evaluated, but rather the total measure of such points. Thus for node 1, $S_3=\{4, 11, 12, 13\}$, would have

$$\theta_3=\mu_4x+(\mu_{11}+\mu_{12}+\mu_{13})x^2.$$

Formula (13) is applied to the several θ_r and integration from 0 to 1 gives the value.

If the Banzhaf-Coleman index, instead, were desired, we would use the formula $\Psi_1[w]=F_1(1/2)$. (See Owen [1975].)

5. Centrality. In [1982], Grofman and Owen suggested that centrality in social networks could be measured by computing the power indices of games such as we have in this paper. In fact, if v is a game, Γ a graph, and $w=v/\Gamma$, then $\Psi_i[w]$ or $\phi_i[w]$ measure something which is related to player i's ability to "bring people together".

Generally speaking, if we consider the quantity $\phi_i[w]$, it will depend on two things: (1) i's role in the original game v; and (2) i's position in the graph Γ. Thus one possibility would be to measure both $\phi_i[w]$ and $\phi_i[v]$; the difference would then measure i's centrality in some way. An alternative would be to choose a very symmetric game v: in this case any difference between the values $\phi_i[w]$ and $\phi_j[w]$ would be purely due to the difference in centrality of nodes i and j. It is this latter approach which was used in Owen and Grofman [1982], with the symmetric game $v(S)=s(s-1)$, as analyzed above, in § 4(d). Readers might like to consider other possible games.

Appendix. Proofs.

We give here proofs of the theorems. The following lemma is of use.

LEMMA. *For integer* $n \geqq 0$,

$$\sum_{k=0}^{n}(-1)^n\binom{n}{k}=\begin{cases}0 & if\, n\geqq 1,\\ 1 & if\, n=0.\end{cases} \tag{15}$$

Proof. This is the binomial expansion of $(1-1)^n$, and is thus 0 if n is positive. For $n=0$, the sum reduces to a single summand, 1.

Proof of Theorem 1. For a fixed Γ, the components $T_1, T_2, \cdots, T_m$ of the set S are fixed. Thus formula (4) is clearly linear.

To prove L_Γ maps Q_N into itself, assume v is super-additive. Let $S \cap T = \varnothing$, set $K = S \cup T$, let S_i and T_j be the components of S and T respectively, and let $K_1, \cdots, K_p$ be the components of K.

Every component of S or T will be a subset of one of the components of K. Now, if K_l is a component of K, we have

$$K_l = \bigcup S_i \cup \bigcup T_j$$

where the unions are taken over some subcollections. By super-additivity of v,

$$v(K_l) \geqq \sum v(S_i) + \sum v(T_j)$$

where the sums are taken over all i, $S_i \subset K_l$, and all j, $T_j \subset K_l$. Then

$$w(K) = \sum_{l=1}^{p} v(K_l) \geqq \sum_{i=1}^{m} v(S_i) + \sum_{j=1}^{q} v(T_j)$$

where the right-hand sums now include all the components of S and T. Hence, $w(K) \geqq w(S) + w(T)$. Thus $w \in Q_N$ as desired.

Proof of Theorem 2. Let S be disconnected, and let T be arbitrary. If T is not a subset of S, or if there is no connected K with $T \subset K \subset S$, then $C_T(S) = 0$, and, moreover, $C_T(M) = 0$ for all $M \subset S$. Thus $\Delta_{C_T(S)} = 0$.

Suppose, however, there is some connected K with $T \subset K \subset S$. Since K is connected, S has a component L such that $K \subset L$.

If $M \subset S$, we write $M = M_1 \cup M_2$, where $M_1 \subset L$, $M_2 \subset S - L$. Clearly $C_T(M) = C_T(M_1)$ whatever M_2 may be. Now,

$$\begin{aligned}
\Delta_{C_T(S)} &= \sum_{M \subset S} (-1)^{s-m} C_T(M) \\
&= \sum_{M_1 \subset L} \sum_{M_2 \subset S-L} (-1)^{l-m_1}(-1)^{s-l-m_2} C_T(M_1) \\
&= \sum_{M_1 \subset L} (-1)^{l-m_1} C_T(M_1) \left[\sum_{M_2 \subset S-L} (-1)^{s-l-m_2} \right] \\
&= \sum_{M_1 \subset L} (-1)^{l-m_1} C_T(M_1) \left[\sum_{M_2=0}^{s-l} (-1)^{s-l-m_2} \binom{s-l}{m_2} \right]
\end{aligned}$$

since $S - L$ will have $\binom{s-l}{m_2}$ subsets of cardinality m_2. Now, since S is disconnected, $S - L \neq \varnothing$, so $s - l \geqq 1$ and hence, by the lemma, this last bracket is zero. Hence $\Delta(S) = 0$.

Proof of Theorem 4. Let Γ be a tree, and let S, T be connected subsets of N. We must show $S \cap T$ is connected.

Take any two nodes, i and j, in $S \cap T$. Since Γ is a tree, there is a unique path between i and j. Since S and T are both connected, this path must be a subgraph of both the partial graphs Γ_S and Γ_T. But then it must be a subgraph of their intersection $\Gamma_S \cap \Gamma_T$, which is the same as $\Gamma_{S \cap T}$. Thus i and j are connected in $\Gamma_{S \cap T}$. It follows $S \cap T$ is connected.

Proof of Theorem 5. *Necessity.* We have already seen $H(T)$ is connected and $T \subset H(T)$. To prove (iii), suppose $j \in E(S)$ but $j \notin T$. In this case $T \subset S - \{j\}$, and $S - \{j\}$ is connected. Thus S is not minimal, and so $S \neq H(T)$.

Sufficiency. Suppose (i), (ii), and (iii) hold. then $T \subset S$ and S is connected; we need to show S is minimal in this sense.

Since Γ is a tree and S is connected, Γ_S is also a tree. If $j \in S$, then either j is extreme or not extreme in S. If j is extreme, then by (iii) $j \in T$ and so $j \in H(T)$.

If j is not extreme in S, then it lies on a path joining extreme nodes i, $k \in S$. (This is most easily seen if we merely start from j in two different directions: eventually the paths must end at extreme points i, k since there are no circuits.) Now, by (iii), i, $k \in T$ and hence any connected set containing T must include the entire path from i to k and, in particular, node j. Thus $j \in H(T)$.

Thus, whether j is extreme or not, $j \in H(T)$. But j was arbitrary in S, and so $S \subset H(T)$. Hence S is minimal, i.e. $S = H(T)$.

Example (c). The expression

$$-\sum_{k=1}^{s} (-1)^{s-k}\binom{s}{k}$$

can be rewritten as

$$-\sum_{k=0}^{s} (-1)^{s-k}\binom{s}{k} + (-1)^s = (-1)^s - (-1)^s\left[\sum_{k=0}^{s} (-1)^{-k}\binom{s}{k}\right].$$

Since $(-1)^{-k} = (-1)^k$, application of the lemma tells us that the term in brackets will vanish whenever $s \geqq 1$, so that only the term $(-1)^s$ will remain. In case $S = \varnothing$, or if S is not a subset of T, $\Delta_p(S)$ is clearly zero.

Similarly, the expression

$$\sum_{k=e}^{s'} (-1)^k\binom{s'-e}{k-e}$$

can be rewritten, setting $j = k - e$, as

$$\sum_{j=0}^{s'-e} (-1)^j(-1)^e\binom{s'-e}{j} = (-1)^e\left[\sum_{j=0}^{s'-e} (-1)^j\binom{s'-e}{j}\right]$$

and, by the lemma, this will be 0 if $s' - e \geqq 1$, and $(-1)^e$ if $s' - e = 0$.

REFERENCES

JOHN F. BANZHAF, III, *Weighted voting doesn't work*, Rutgers Law Review, 19 (1965), pp. 317-343.

JAMES COLEMAN, *Control of collectivities and the power of a collectivity to act*, in Social Choice, B. Lieberman, ed., New York, Gordon and Breach, 1971.

ROGER MYERSON, *Graphs and cooperation in games*, Math. Oper. Res., 2 (1977), pp. 225-229.

GUILLERMO OWEN, *Multilinear extensions of games*, Management Sci., 18 (1972), Supplementary issue, P64-P79.

———, *Multilinear extensions and the Banzhaf value*, Nav. Res. Log. Quart., 2 (1975), pp. 741-752.

GUILLERMO OWEN AND BERNARD GROFMAN, *A game-theoretic approach to measuring centrality*, Social Networks, 3 (1982), pp. 213-224.

LLOYD S. SHAPLEY, *A value for n-person games*, in Contributions to the Theory of Games II, A. W. Tucker and H. W. Kuhn, eds., Annals of Math. Study 28, Princeton Univ. Press, Princeton, NJ, 1953, pp. 307-317.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

007

THE CYCLIC COLORING PROBLEM AND ESTIMATION OF SPARSE HESSIAN MATRICES*

THOMAS F. COLEMAN† AND JIN-YI CAI†

Abstract. Numerical optimization algorithms often require the (symmetric) matrix of second derivatives, $\nabla^2 f(x)$. If the Hessian matrix is large and sparse, then estimation by finite differences can be quite attractive since several schemes allow for estimation in much fewer than n gradient evaluations.

The purpose of this paper is to analyze, from a combinatorial point of view, a class of methods known as substitution methods. We present a concise characterization of such methods in graph-theoretic terms. Using this characterization, we develop a complexity analysis of the general problem and derive a roundoff error bound on the Hessian approximation. Moreover, the graph model immediately reveals procedures to effect the substitution process optimally (i.e. using fewest possible substitutions given the differencing directions) in space proportional to the number of nonzeros in the Hessian matrix.

Key words. graph coloring, estimation of Hessian matrices, sparsity, differentiation, numerical differences, NP-complete problems, unconstrained minimization

AMS(MOS) subject classifications. 65K05, 65K10, 65H10, 68L10

1. Introduction.

We are concerned with the estimation of a large sparse symmetric matrix of second derivatives $\nabla^2 f(x)$ for some problem function $f: R^n \to R^1$. In particular, we note that the product $\nabla^2 f(x) \cdot d$ can be estimated, for example, by forward differences

$$\nabla^2 f(x) \cdot d = [\nabla f(x+d) - \nabla f(x)] + o(\|d\|). \tag{1.1}$$

When the structure of $\nabla^2 f(x)$ is known, then usually a few well chosen differencing directions $d_1, \cdots, d_p$ affords the recovery of estimates of all nonzeros of $\nabla^2 f(x)$. Let us denote our estimate by H. We will assume that the sparsity pattern of H is known; the diagonal elements are specified as nonzero; H is symmetric. (Restricting the diagonal to be zero-free is reasonable in many contexts: In particular, a minimizer of f usually possesses a positive definite Hessian matrix.) We will be concerned with methods that use differencing directions $d_1, d_2, \cdots, d_p$ that are based on a *partition* of columns $C_1, \cdots, C_p$. In particular, let S_k denote the set of columns in group C_k and let h_i be the steplength associated with column i, $i = 1, \cdots, n$. Finally, define

$$d_k = \sum_{i \in S_k} h_i e_i \tag{1.2}$$

for $k = 1, \cdots, p$, where e_i is the ith column of the identity.

There has been considerable work recently concerned with this problem, especially with trying to make p as small as possible. Curtis, Powell, and Reid [1974] suggested a method, *CPR*, for the unsymmetric problem. Their idea was to build groups of *structurally independent* columns in a left-to-right greedy fashion. (Two columns (vectors) x, y are *structurally independent* if $x_i * y_i = 0$, for all i.) It is easy to see that such a p-partition allows for the estimation of a matrix with p differencing directions. Specifically, let $C_1, \cdots, C_p$ be a partition of the columns of H where each group consists of structurally independent columns. Then, if $[\nabla f(x+d_k) - \nabla f(x)]_i \neq 0$ it follows that there is exactly one column j in group C_k with H_{ij} a designated nonzero

* Received by the editors October 10, 1984, and in revised form April 24, 1985. This work 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-AC02-83ER13069.

† Computer Science Department, Cornell University, Ithaca, New York 14853.

and we can assign

$$H_{ij} \leftarrow \frac{[\nabla f(x+d_k)-\nabla f(x)]_i}{h_j}.$$

Coleman and Moré [1983] analyzed and modified this method by taking a combinatorial point of view. In particular, a *column intersection graph* can be formed by associating with each column i of H a node v_i and defining an edge between node v_i and node v_j iff there is an index k such that both H_{ki} and H_{kj} are nonzeros. A p-coloring of this graph is an assignment, ϕ, of "colors" to nodes such that if there is an edge between node v_i and node v_j then $\phi(v_i) \neq \phi(v_j)$. It is not hard to see that a p-coloring of this graph induces a valid partition of structurally independent columns and vice versa.

Coleman, Garbow, and Moré [1984] have developed FORTRAN 77 codes based on this work. Such (unsymmetric) methods can be applied to the symmetric problem (McCormick [1983] discusses the complexity of this approach); however it is probably worthwhile using symmetry when it is present.

Powell and Toint [1979] were the first to try to exploit symmetry. They pointed out that symmetry can be used both in a direct and an indirect fashion. A direct method is one in which each unknown of H is determined independently of the others. More specifically, let $C_1, \cdots, C_p$ be a partition of the columns of H. Since each off-diagonal nonzero is represented twice, it is no longer necessary that each group consist of structurally independent columns. It is necessary, however, that for each nonzero (i, j) *either* column i resides in a group C_r such that no other column in this group has a nonzero in row j *or* column j resides in a group C_s such that no other column in this group has a nonzero in row i. If the latter condition were true then H_{ij} would be determined

$$H_{ij} \leftarrow \frac{[\nabla f(x+d_s)-\nabla f(x)]_i}{h_j}$$

and $H_{ji} \leftarrow H_{ij}$. Clearly a similar (symmetric) rule would hold for the former condition.

Coleman and Moré [1984] analyzed such methods from a combinatorial point of view and produced a simple graph-theoretic characterization of all partitions that can be used to induce a direct symmetry-exploiting determination of H. Let us represent the structure of H by the usual adjacency graph $G(H)=(V(H), E(H))$. That is, if H is a symmetric matrix of order n, then $V(H)$ consists of n vertices $v_1, \cdots, v_n$ (associate column i of H with vertex v_i) and $E(H)$ consists of pairs of vertices (edges) where $(v_i, v_j) \in E(H)$ if and only if $H_{ij}(H_{ji})$ is considered a nonzero. A p-partition of the columns of H, $C_1, \cdots, C_p$ can be viewed as an assignment of colors, ϕ, to the nodes of G, $\phi\colon V \to \{1, \cdots, p\}$. This assignment is a *p-coloring* if $(v, w) \in E \Rightarrow \phi(v) \neq \phi(w)$. A *path p-coloring* is a p-coloring with the additional stipulation that every path in G of length 4 (distinct) vertices uses at least 3 colors. The characterization of direct symmetric methods given by Coleman and Moré is simply

THEOREM 1.1. *The mapping ϕ is a path p-coloring if and only if ϕ induces a partition of the columns of H consistent with direct determination.*

Note: We have changed the notation used by Coleman and Moré [1984]; here we use "path coloring" instead of "symmetric coloring" because in our context the term path coloring is more appropriate.

This characterization led to a deeper understanding of the direct estimation problem on symmetric structures which in turn yielded a complexity analysis and algorithmic possibilities.

Indirect estimation of symmetric matrices may be preferable because fewer groups (i.e. differencing directions) will be needed, in general. Powell and Toint concentrated on substitution methods where directions are chosen so that nonzeros can be determined via a substitution process. (They restricted their attention, as we do, to substitution methods based on a partition of columns.) So in this case there is interdependence of the matrix unknowns (nonzeros) to the degree that an underlying lower triangular system is defined. Powell and Toint proposed an algorithm to determine the differencing directions and then solve for the unknowns (lower triangular substitution method (LTS)). Subsequently, Coleman and Moré [1984] analyzed this process from a combinatorial point of view. This analysis led to a modified and empirically superior procedure (the resulting FORTRAN 77 code is described in Coleman, Garbow, and Moré [1985]). However, a simple insightful characterization, in the vein of Theorem 1.1, was not provided.

The purpose of this paper is to provide such a characterization. This result is as simple as Theorem 1.1 and is clearly the analogous result. This view provides enormous insight into the combinatorial nature of the problem as well as suggesting algorithmic possibilities. Furthermore, the graph theoretic interpretation reveals that if a partition of columns allows for the recovery of H via a substitution process, then it is always possible to do so efficiently. In particular, every unknown can be solved for in (roughly) less than $n/2$ substitutions and the space required to compute H is proportional to the number of nonzeros. This is somewhat surprising since the Powell-Toint procedure relies heavily on a regular matrix structure produced by LTS which is not present for an arbitrary feasible partition. Finally, the graph model allows one to derive a growth of error bound for a general substitution method, which is essentially analogous to the result achieved by Powell and Toint for a specific method, LTS.

Section 2 will provide the characterization of substitution methods followed by a roundoff error discussion in § 3. In § 4 we establish the complexity of the problem and discuss its combinatorial relationship to the symmetric direct problem (path coloring). Section 5 deals with algorithms for effecting the substitution process in space proportional to $|E|$ (i.e. the number of nonzeros of H). Finally, observations on parallelism are provided in § 6.

2. Substitution methods and cyclic coloring. A partition of columns of a symmetric matrix induces a substitution method if there is an ordering of the matrix unknowns such that all unknowns can be solved for, in that order, using symmetry and previously solved elements. This notion is fully general (subject to the partition restriction) but seems to be a difficult one to work with. There is, however, a very elegant and simple graph theoretic interpretation. The major purpose of this section is to present this characterization.

First it is necessary to formalize the concept of a substitution method in matrix terms. Let U be the set of indices of matrix unknowns (identify (i, j) with (j, i)) and suppose that U is ordered: $U=\{(i_k, j_k)\}$. Let the columns of H be partitioned $\{C_1, \cdots, C_p\}$ and define

$$(*) \qquad S_0=\varnothing, \qquad S_k=S_{k-1}\cup\{(i_k, j_k)\}, \quad 1\leqq k\leqq|U|.$$

The ordering *induces a substitution method* iff

either j_k belongs to a group C, say, and if l is any other column in C with a nonzero in row i_k then $(i_k, l)\in S_{k-1}$ *or* i_k belongs to a group C', say, and if l' is another column in C' with a nonzero in row j_k then $(j_k, l')\in S_{k-1}$.

The essence of this statement is that, at the kth step, it is possible to solve for element (i_k, j_k) or, equivalently (j_k, i_k), by substitution. We call a partition, for which there exists such an ordering, *substitutable.* For example, if H is a tridiagonal matrix, then it is easy to verify that the partition $(\{1,3,\cdots\}, \{2,4,\cdots\})$ is substitutable.

Obviously there are substitutable partitions for any symmetric matrix. For example, every partition consistent with a path coloring is substitutable. Alternatively, a partition that induces a "lower triangular substitution method" is substitutable. (Coleman and Moré [1984] and Powell and Toint [1979] discussed lower triangular substitution methods.) However, here we are interested in minimizing the number of groups in a general substitutable partition. The above 2 examples are restrictive in that they consider only particular classes of substitutable partitions. The general problem is

Partition problem. Obtain a substitutable partition of the columns of a given symmetric marix H with the fewest groups.

How difficult is the partition problem? This is a hard question to answer considering the rather clumsy matrix formalization of a substitution method. Fortunately a substitutable partition has a simple expression in the language of graphs.

DEFINITION. A mapping $\phi: V \to \{1,2,\cdots,p\}$ is a *cyclic p-coloring* of G if ϕ is a p-coloring and if ϕ uses at least 3 colors in every cycle of G.

As the following theorem indicates, we now have a simple characterization of a substitutable partition.

THEOREM 2.1. *Let H be a symmetric matrix with a nonzero diagonal. The mapping ϕ induces a substitution method if and only if ϕ is a cyclic coloring of $G(H)$.*

Before providing the proof, let us consider an informal argument based on the following example. Let the adjacency graph of H, $G(H)$ be as shown in Fig. 1. Both assignments of the colors r, s, t are valid colorings but assignment 1 is *not a valid cyclic coloring*: the cycle v_1, v_2, v_3, v_4 uses only 2 colors. Assignment 2 is a valid cyclic coloring. The edges (off-diagonal nonzeros) can be determined by considering each pair of colors in turn. For example, consider the subgraph, $F_{r,s}$, induced by the nodes colored r or s as shown in Fig. 2. Edges $(1,8)$, $(3,7)$, and $(3,9)$ can all be determined immediately. Consider for example edge $(3,7)$. Column 3 has a nonzero in row 7 and residues in group C_r. There is no other column in group C_r with a nonzero in row 7 (else node 7 would have another incident r-node). Therefore, $H_{7,3}$ (hence $H_{3,7}$) can be determined directly. Once $(3,7)$ and $(3,9)$ are determined, edge $(2,3)$ can be computed: column 2 has a nonzero in row 3 and resides in group C_s. Columns 7 and 9 are the other columns in group C_s with nonzeros in row 3. However, $H_{3,7}$ and $H_{3,9}$ are now known quantities; hence, $H_{3,2}$ can be computed with 2 substitutions.

Clearly the process can be carried to completion until every edge in $F_{r,s}$ is determined. (It is easy to see that the diagonal elements can be directly determined: this follows from the fact that ϕ is a coloring.) But every pair of colors induces a forest, otherwise ϕ would not be a cyclic coloring, and therefore every nonzero can be determined by considering each pair of colors in turn.

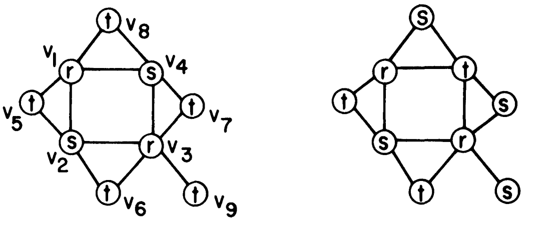


FIG. 1

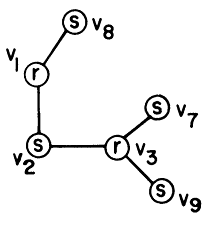


FIG. 2

The "if" part of Theorem 2.1 is proved along the lines of the example given above. The "only if" part is perhaps a bit surprising but not difficult to prove.

Proof of Theorem 2.1. First we prove that every cyclic coloring of $G(H)$ induces a substitution method. Since ϕ is a coloring, every diagonal element of H can be computed. In particular, if $\phi(v_j)=r$, then j is the only column in group C_r with a nonzero in row j (otherwise ϕ is not a coloring). But then (1.1) and (1.2) yield

$$H_{jj}=\frac{[\nabla f(x+d_r)-\nabla f(x)]_j}{h_j}.$$

Consider next $(i,j)\in U$, $i\neq j$. Suppose that column i is in group C_r and column j is in group C_s. Clearly, since ϕ is a coloring, $r\neq s$. Consider the subgraph induced by the nodes colored r and the nodes colored s, say $F_{r,s}$. Since ϕ is a cyclic coloring, $F_{r,s}$ contains no cycle and therefore is a forest. The edges in $F_{r,s}$ correspond to off-diagonal unknowns of H. They can be solved, or ordered, independent of the rest of the unknowns of H. In particular, each leaf-incident edge can be solved directly since there is no conflict. We can now "delete" all such edges and consider each new leaf-incident edge. Each such edge is now incident to known edges and can therefore be solved. Clearly the process can be repeated until an edge-less graph remains. The entire procedure can now be repeated for each pair of colors until every unknown is determined.

We now show that if ϕ induces a substitution method, then ϕ is a cyclic coloring. First it is clear that ϕ must be a valid coloring, otherwise the diagonal elements would not be determined. To see this, suppose that $(i,j)\in U$, and v_i and v_j are assigned the same color r. Hence both column j and column i are in the same group, C_r. Since column j belongs only to group C_r, it follows from (*) that H_{jj} can be determined only after either H_{ij} or H_{ji} is determined. Similarly, H_{ii} can be determined only after either H_{ij} or H_{ji} is determined. But the determination of one of H_{ij}, H_{ji} must be preceded by the determination of one of H_{ii}, H_{jj}, by (*), which is a contradiction.

Suppose then that ϕ is a coloring but is not a cyclic coloring. Hence there must be a cycle, with at least 4 edges, colored with just 2 colors, say r, s. Let (i,j) be the *first* edge in this cycle to be solved (ordered) and let us assume, without loss of generality, that v_i is colored r and v_j is colored s. Let node v_i be incident also to node v_h(on the cycle) and let v_j be incident also to node v_k (on the cycle), as illustrated in Fig. 3. But (i,j) cannot be determined from group C_s because columns j and h both reside in this group with nonzeros in row i (and (i,h) is not yet known (ordered)). Similarly, (j,i) is not determined from group C_r because columns i and k both reside in this group with nonzeros in row j (and (j,k) is not yet known (ordered)). Therefore no edge in this cycle can be solved first and ϕ cannot induce a substitution method. □

Hence the partition problem is equivalent to the

Cyclic coloring problem: Obtain a minimum cyclic coloring of $G(H)$.

Note that once we have found a cyclic coloring of $G(H)$, then the coloring induces a substitutable partition and the corresponding ordering of U is available, as the proof of Theorem 2.1 indicates. A tridiagonal matrix provides a simple example. The graph

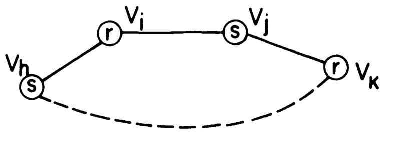


FIG. 3

is shown in Fig. 4 and a valid cyclic coloring is provided by assigning r to the even nodes and s to the odd nodes. The diagonal elements can be solved directly and the off-diagonal elements are obtained via substitution: edges $(1,2)$ and $(n-1,n)$ are obtained first (directly), followed by $(2,3)$ and $(n-2,n-1)$, with 1 substitution each, and so on. The middle edge will be the last determined element with approximately $\frac{1}{2}n$ dependencies or substitutions.

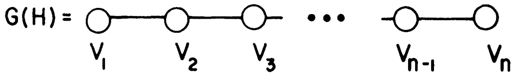


FIG. 4

Suppose we modify the above example by adding an edge from node v_1 to node v_n. A cyclic coloring would then require 3 colors; for example, we could use our previous assignment except we apply a new color, t, to node 1. Now $(1,2)$ and $(1,n)$ can be determined directly (or, ordered first) and the remaining elements can be determined, as before, via substitution.

3. Substitution methods and roundoff error. The above two examples raise an interesting question with numerical significance: Is there a limit to the number of dependencies or substitutions? The amount of computational work as well as the potential growth of roundoff error depends, in part, on this number; therefore, a bound tighter than the total number of nonzeros in H may be consequential. Powell and Toint [1979] established that for a particular class of substitution methods, triangular substitution methods, the bound is $n-2$. The cyclic coloring characterization leads us immediately to a more general result. Every unknown can be determined by considering the forest induced by a particular pair of colors. But each forest can have at most $n-1$ edges and therefore we have the following result.

THEOREM 3.1. *Let ϕ be a substitutable partition. Then, each unknown in H is dependent on at most $n-2$ other unknowns.*

Clearly this result is the best possible worst case upper bound, if we allow any possible feasible ordering of the unknowns or edges. To see this, just consider the tridiagonal case: if the edges are solved from one end of $G(H)$ to the other, then the last edge requires $n-2$ substitutions. However, certain orderings are preferable over others. For example, in the tridiagonal case one can achieve a bound of $\lfloor\frac{1}{2}(n-2)\rfloor$ if each edge is solved by substituting from the nearest end of $G(H)$. It is not hard to see that, over different orderings, this is the best possible worst case upper bound; again, just consider the tridiagonal case.

Is it possible to order the unknowns, in general, so that the maximum number of substitutions is less than or equal to $\lfloor\frac{1}{2}(n-2)\rfloor$? In order to answer this question, consider when it is feasible, during the solution process, to solve for edge $l \triangleq (x,y)$ in $T_{r,s}$ where $T_{r,s}$ is a tree in the forest induced by the colors r and s. Note that if an edge (x,y) is removed from a tree (but the nodes x, y are not removed) then two subtrees remain: Let

$$T^l_{r,s}(x) = (V^l_{r,s}(x), E^l_{r,s}(x)), \qquad T^l_{r,s}(y) = (V^l_{r,s}(y), E^l_{r,s}(y))$$

represent the two subtrees, rooted at x and y respectively, that remain when edge l is removed from $T_{r,s}$: consider Fig. 5. It is clear that (x,y) is ready to be solved *if and only if* either every edge in $T^l_{r,s}(x)$ is solved or every edge in $T^l_{r,s}(y)$ is solved. Furthermore, (x,y) requires at least

$$\textit{mincost}\ (x,y) \triangleq \min\{|E^l_{r,s}(x)|, |E^l_{r,s}(y)|\}$$

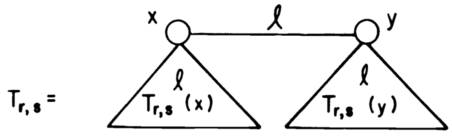


FIG. 5

substitutions. Note that an ordering that computes edge (x, y) using *mincost* (x, y) substitutions, for each edge (x, y), is optimal and requires less than $\lfloor\frac{1}{2}(n-2)\rfloor$ substitutions for each edge.

Such an ordering is possible and is provided by the following algorithm. Let $T \triangleq (V_T, E_T)$ be the tree under consideration, with $|V| = n_T \leqq n$.

```
ALGORITHM solve_tree
  $T_1 \triangleq (V_1, E_1)$ where $V_1 = V_T$, $E_1 = E_T$
  for each vertex $v \in V_1$ do value$(v) \leftarrow 0$ endo
  for $i = 1$ to $|E_1|$ do
    choose a leaf $x_i$ of $T_i$, of smallest value
    let $y_i$ be the vertex such that $(x_i, y_i) \in E_i$
    value $(y_i) \leftarrow$ value $(y_i)$ + value $(x_i) + 1$
    solve $(x_i, y_i)$
    $E_{i+1} \leftarrow E_i - \{(x_i, y_i)\}$
    $V_{i+1} \leftarrow V_i - \{x_i\}$
    $T_{i+1} \leftarrow (V_{i+1}, E_{i+1})$
  endo
end solve_tree
```

THEOREM 3.2. *Algorithm solve_tree solves for each edge $(x, y) \in E_T$ using the fewest possible substitutions, mincost (x, y). Hence, solve_tree requires at most*

$$\max\{\textit{mincost}\,(x, y) : (x, y) \in E\} \leqq \lfloor\tfrac{1}{2}(n_T-2)\rfloor \leqq \lfloor\tfrac{1}{2}(n-2)\rfloor$$

substitutions to determine any edge of E_T.

Proof. First we establish that algorithm *solve_tree* terminates: Since T_i is a tree and x_i is a leaf in T_i, it follows that T_{i+1} is a tree. Therefore T_{i+1} will have a leaf (indeed at least 2) and x_{i+1} will be found. It follows that the algorithm will determine every edge.

Assume then that *solve_tree* does not solve for each edge using the fewest possible substitutions. In particular suppose that at step i vertex x_i is the chosen leaf in T_i and edge $l \triangleq (x_i, y_i)$ will be determined nonoptimally. That is, $|E^l(x_i)| > |E^l(y_i)|$ and hence $|E^l(x_i)| = \textit{value}\,(x_i) > \lfloor\frac{1}{2}(n_T-2)\rfloor$, since $|E^l(x_i)| + |E^l(y_i)| = n_T - 2$.

Consider a leaf, v_i, in $T^l(y_i)$ (there must be at least 1). But

$$|E^l(x_i)| > \lfloor\tfrac{1}{2}(n_T-2)\rfloor \Rightarrow |E^l(y_i)| \leqq \lfloor\tfrac{1}{2}(n_T-2)\rfloor$$

and therefore,

$$\textit{value}\,(v_i) \leqq \lfloor\tfrac{1}{2}(n_T-2)\rfloor < \textit{value}\,(x_i).$$

Therefore (x_i, y_i) would not be chosen at step i, a contradiction. □

In summary, Theorem 3.2 says that for an arbitrary substitutable partition algorithm *solve_tree* will compute each edge with the fewest possible substitutions (with respect to that partition) and that number is always bounded by $\lfloor\frac{1}{2}(n-2)\rfloor$.

We conclude § 3 by considering the accuracy of the estimated Hessian matrix in more detail. We will show that an error bound, similar to that achieved by Powell and

Toint [1979] for a particular class of algorithms (lower triangular substitution methods) holds for any substitution method provided the unknowns are solved for in the manner suggested by *solve_tree.*

Every substitutable partition with p groups, or cyclic p-coloring, allows for the recovery of the matrix unknowns via a back substitution process provided the differencing vectors are consistent with the coloring ϕ. In particular, let S_k denote the set of nodes (columns) colored k (i.e. in C_k) and again define

$$d_k = \sum_{i \in S_k} h_i e_i$$

where h_i is the step-length associated with column i. Let x be a given point in R^n and define $u_k = \nabla f(x+d_k) - \nabla f(x)$, for $k = 1, \cdots, p$. If H denotes the approximation to $\nabla^2 f(x)$, then, since ϕ is a coloring and by (1.2),

$$H_{jj} \cdot h_j = \nabla f(x+d_k)_j - \nabla f(x)_j \quad \text{for } j \in S_k$$

and therefore, since every column belongs to a group, every diagonal element can be determined. The diagonal approximations are defined by these equations, and will not participate in any subsequent calculations. Indeed such equations usually guide the choice of h_j: h_j is chosen to balance truncation and roundoff errors in order to approximate the diagonal elements as accurately as possible (e.g. Gill, Murray, Saunders, and Wright [1983]).

Our previous analysis has shown that it is only necessary to consider 2 colors (directions) at a time when solving for the off-diagonal elements. Let us concern ourselves then with a tree, $T_{r,s}$, induced by colors r and s. Let $u_r = \nabla f(x+d_r) - \nabla f(x)$, $u_s = \nabla f(x+d_s) - \nabla f(x)$ and let

$$\hat{u}_r = u_r + \varepsilon_r, \qquad \hat{u}_s = u_s + \varepsilon_s$$

denote the computed quantities (i.e. contaminated with rounding error).

The solution process is provided by algorithm *solve_tree* with the statement "*solve* (x_i, y_i)" expanded, to read

$$H_{ij} \cdot h_j = (\hat{u}_c)_i - \sum_{k \in N(i)} H_{ik} \cdot h_k \tag{3.1}$$

where we *identify vertex* x_i *with index* i, *and vertex* y_i *with index* j. $N(i)$ is the set of neighbours of node x_i in $T_{r,s}(x_i)$, and $c = \phi(y_i)$, which is one of r, s (for brevity we will write $T_{r,s}(x_i)$ instead of $T^{l_i}_{r,s}(x_i)$ where $l_i = (x_i, y_i)$. In other words, when H_{ij} is solved for, every other element in row i of columns in group C_c has already been solved for; the right-hand side of (3.1) is adjusted accordingly.

Following Powell and Toint, we define the error matrix F to be $H - \nabla^2 f$ and let

$$(\delta_c)_i = (\hat{u}_c)_i - \sum_{k \in N(i) \cup \{j\}} (\nabla^2 f(x))_{ik} \cdot h_k. \tag{3.2}$$

In other words, $(\delta_c)_i$ measures the difference between the computed quantity $(\hat{u}_c)_i$ and the ideal $(\nabla^2 f(x) \cdot d_c)_i$. Hence $(\delta_c)_i$ is a composite of roundoff and truncation errors. If we assume that the second derivatives of f are Lipschitz continuous, then a standard bound is obtained:

$$\eta \triangleq \max_{c,i} \{|(\delta_c)_i|\} \leqq C \cdot \max_k \{|h_k|^2\} + \max_{c,i} \{|(\varepsilon_c)_i|\}$$

where C is a positive constant.

The following result establishes a bound on the elements in the error matrix F.

THEOREM 3.3. *If H is obtained by algorithm solve_tree (with "solve (x_i, y_i)" effected by* 3.1) *then*

$$|F_{ij}| \leqq (|E_{r,s}(x_i)|+1) \cdot \eta \cdot \max_{i,j,k}\left\{\frac{|h_k|}{|h_ih_j|}\right\}$$

$$\leqq (\lfloor \tfrac{1}{2}n \rfloor) \cdot \eta \cdot \max_{i,j,k}\left\{\frac{|h_k|}{|h_ih_j|}\right\}$$

where again we identify column i with node x_i, column j with node y_i, and $\{\phi(x_i), \phi(y_i)\} = \{r, s\}$.

Proof. Combining (3.1), (3.2) and the definition of F yields

$$(\delta_c)_i = F_{ij} \cdot h_j + \sum_{k \in N(i)} F_{ik} \cdot h_k$$

which implies the bound

$$|F_{ij}h_ih_j| \leqq |(\delta_c)_i \cdot h_i| + \sum_{k \in N(i)} |h_ih_kF_{ik}|$$

$$= |h_i(\delta_c)_i| + \sum_{k \in N(i)} |F_{ki}h_kh_i|.$$

But this same decomposition can be applied, recursively, to each $F_{ki}h_kh_i$, for $k \in N(i)$, to yield

$$|F_{ij}h_ih_j| \leqq \sum_{w \in V_{r,s}(x_i)} |(\delta_c)_w \cdot h_w| \tag{3.3}$$

where $T_{r,s}(x_i) = (V_{r,s}(x_i), E_{r,s}(x_i))$. Since the tree $T_{r,s}(x_i)$ has $(|E_{r,s}(x_i)|+1)$ nodes, the result follows immediately from (3.3) and Theorem 3.2. □

One can conclude from this result that the growth of roundoff error is quite limited if the steplength does not vary greatly in size. On the other hand, if there is significant variance (recall that stepsizes are chosen to accurately approximate diagonal elements) then this result may allow for unacceptable growth of error: a direct method may be preferable.

Indeed recent experiments by Coleman, Garbow, and Moré [1985] support the conclusion that unacceptable error pollution can occur when stepsizes vary noticeably in size. The resulting matrix is essentially unuseable as an approximation to the Hessian. This suggests that an automatic monitoring process that switches from an indirect method to a direct method, when necessary, might be useful. Unfortunately we have no specific suggestions at the moment as to what quantities to monitor. (Of course it is always possible to estimate the Hessian by both an indirect and a direct method, occasionally, and compare the resulting matrices.)

4. The cyclic chromatic number. How difficult is the cyclic coloring problem? We address this question in this section. The reader who is unfamiliar with the fundamentals of complexity theory and NP-completeness is urged to consult the excellent resource book *Computers and Intractability: A Guide to the Theory of* NP-*Completeness*, by Michael R. Garey and David S. Johnson [1979].

We will first consider the cyclic coloring decision problem (CCDP) and show that this problem is NP-complete: we do this by transforming the general graph coloring decision problem (CDP). We then conclude that the corresponding optimizaion problem, the cyclic coloring problem, is NP-hard. The consequence of this result is just this: if we could solve the cyclic coloring problem in polynomial time (P-time) then we could also solve the graph coloring problem in P-time (as well as a host of other "intractable" problems). Since this is deemed highly unlikely, an expedient approach to our problem is to investigate efficient heuristic and approximation schemes (we discuss this in § 5).

It is common, when considering complexity questions related to discrete optimization problems, to consider the decision problem formulation. In this case we have the

Cyclic coloring decision problem (CCDP): Given an integer $p \geqq 3$ and an arbitrary graph G, is it possible to assign a cyclic p-coloring to the nodes of G?

We have excluded the simple cases $p = 1, 2$ since it is easy to see that polynomial algorithms exist for such cases. The following theorem shows that CCDP is not so simple for $p \geqq 3$.

THEOREM 4.1. *CCDP is* NP-*complete.*

Proof. The first step is to show that CCDP is in the class NP. In particular, we must show that we can validate, in P-time, whether or not a particular assignment of p colors is indeed a cyclic coloring. To do this, one must merely consider each pair of colors, in turn, and decide whether or not the induced graph is a forest. Clearly this is a polynomial time operation.

We now proceed to transform the general coloring problem (CDP), which is known to be NP-complete, to CCDP. Consider an arbitrary graph $G = (V, E)$ and integer $p \geqq 3$. Let $|V| = n$ and $|E| = m$. We construct a new graph, $G' = (V', E')$ as follows. For each edge $e_l = (v_i, v_j) \in E$, define a bipartite graph G'_l with vertices

$$\{v_i, v_j, w_1^{(l)}, \cdots, w_p^{(l)}\}$$

and edges

$$(v_i, w_k^{(l)}), \qquad (v_j, w_k^{(l)}), \qquad k = 1, \cdots, p.$$

Graphically, this transformation is shown in Fig. 6.

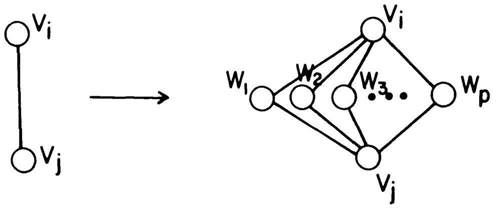


FIG. 6

Now define a bipartite graph G' by setting

$$V' = V(G) \cup \{w_k^{(l)} \colon 1 \leqq k \leqq p, 1 \leqq l \leqq m\}$$

and

$$E' = \bigcup_{l=1}^{m} E(G'_l).$$

We now show that if G can be p-colored using an assignment ϕ, then G' can be assigned a cyclic p-coloring, ϕ'. In particular, for each $v \in V$ let $\phi'(v) = \phi(v)$. Hence if we consider any G'_l, induced by $e_l = (v_i, v_j) \in E$, then $\phi'(v_i) \neq \phi'(v_j)$. Let ϕ' assign vertices $w_j^l, j = 1, \cdots, p$ any color different from $\phi'(v_i)$ and $\phi'(v_j)$.

We claim that ϕ' is a cyclic p-coloring of G': Clearly any cycle in G' must contain a path $(v_i, w_k^{(l)}, v_j)$ for some $1 \leqq l \leqq m$ and $1 \leqq k \leqq p$ where $(v_i, v_j) \in E$. But ϕ' assigns 3 colors to each such path and hence every cycle uses at least 3 colors. Moreover, the transformation from G to G' can obviously be done in P-time.

Finally we show that if G' can be assigned a cyclic p-coloring, then G can be p-colored. Assume that ϕ' is a cyclic coloring of G'. Define

$$\phi\colon \phi(v_i)=\phi'(v_i), \qquad 1\leqq i\leqq n.$$

We claim that ϕ is a p-coloring of G. Suppose instead that $\phi(v_i)=\phi(v_j)$ where $e_l=(v_i, v_j)\in E$. Then ϕ' must assign a different color to each w_k^l, $1\leqq k\leqq p$; otherwise, there is a bi-colored cycle in G_l'. But it follows that ϕ' uses at least $p+1$ colors, a contradiction. □

The proof above has actually established a stronger result than indicated by the statement of Theorem 4.1, since the constructed graph G' is bipartite.

COROLLARY 4.2. *The cyclic coloring decision problem on bipartite graphs is* NP-*complete.*

Since the *cyclic coloring problem* is the optimization version of CCDP, it follows that it cannot be an easier problem. Hence the cyclic coloring problem is NP-hard (even if we restrict our attention to bipartite graphs).

There is a marked similarity between the proof given above and the NP-completeness proof provided by Coleman and Moré [1984] with respect to the path coloring problem (symmetric *direct* problem). Indeed it turns out that the transformation given above will also establish that the path coloring decision problem is NP-complete. (Recall: $\phi: V\rightarrow\{1, 2, \cdots, p\}$ is a *path p-coloring* of a graph G if ϕ is a p-coloring and if ϕ is not a 2-coloring for any path in G of length 3 edges).

We conclude this section with a short discussion on the relationship between path colorings and cyclic colorings. Let $\chi(G)$, $\chi_\pi(G)$, $\chi_0(G)$ denote the chromatic number, the path chromatic number, and the cyclic chromatic number of graph G, respectively. That is, $\chi(G)$ is the smallest integer p such that G has a p-coloring. Similarly, $\chi_\pi(G)$ $[\chi_0(G)]$ is the smallest integer p such that G has a path p-coloring [cyclic p-coloring].

The first observation is that a path coloring is a cyclic coloring. To see this, consider any cycle O in G and suppose that ϕ is a path coloring. Clearly if O has only three edges then, since ϕ is a coloring, O must be assigned 3 colors. If O has more than 3 edges, then O contains a path connecting 4 distinct vertices and hence at least 3 colors are assigned by ϕ. Therefore,

$$\chi_0(G)\leqq\chi_\pi(G) \tag{4.1}$$

for any graph G.

Of course a cyclic coloring is not necessarily a path coloring: a cycle O of arbitrary large circumference needs only 1 vertex to be assigned a third color and effect a valid cyclic 3-coloring however this assignment is not a valid path coloring in general. This raises an interesting question: how large can $\chi_\pi(G)/\chi_0(G)$ be? This ratio can be arbitrarily close to 2 for band graphs; however we have been unable to prove (or disprove) that this is an upper bound. It seems reasonable to hypothesize that 2 is an upper bound because a band graph G is, in a certain sense, the worst possible graph for path coloring and the best possible graph for cyclic coloring. Specifically, the first and the last inequalities become equalities in (4.2) below for all band graphs sufficiently large. (It is easy to verify the first equality, and Coleman and Moré [1984] proved the latter.)

Finally, since every cyclic coloring of G is a coloring of G, and every coloring of G^2 is a path coloring of G, we can stretch both ends of (4.1) to get

$$\chi(G)\leqq\chi_0(G)\leqq\chi_\pi(G)\leqq\chi(G^2). \tag{4.2}$$

Note that a partition that induces a direct method that ignores symmetry is equivalent to a coloring of G^2 and has at least $\chi(G^2)$ groups (Coleman and Moré [1983]); a partition that induces a direct method that uses symmetry is equivalent to a path coloring of G and has at least $\chi_\pi(G)$ groups; a partition that induces a substitution method is equivalent to a cyclic coloring and has at least $\chi_0(G)$ groups. One final comment on (4.2): each inequality can be made strict by choosing appropriate graphs.

5. Algorithms. The NP-completeness result of the previous section indicates that an efficient heuristic, or approximation scheme, is required. In particular, since it is not crucial that the absolute *fewest* groups be found (though it is desirable), we are willing to settle for an efficient procedure that produces near optimal results in practise. Indeed, such procedures have been suggested by Powell and Toint [1979] and Coleman and Moré [1984]. Furthermore, Coleman and Moré report extensive experimental results. In this section we will interpret such procedures in the light of the new characterization described in this paper. In addition, we will discuss an important computational concern: Given a substitutable partition, is it possible to recover the matrix unknowns (i.e. solve for the edges) in an amount of space proportional to the number of matrix unknowns (i.e. the number of edges)?

We wish to obtain a cyclic coloring of $G(A)$ using few colors. Since efficient heuristic approaches to the ordinary graph coloring problem (i.e. no cyclic restriction) are available, a natural approach is to transform our problem to a general graph coloring problem. In particular, consider adding edges to the given graph $G=(V, E)$ to obtain a completed graph $\bar{G}=(V, \bar{E})$ such that a coloring of $\bar{G}$ is a cyclic coloring of G. Consider the following

```
ALGORITHM add_edge
  let $\pi$: $V \to \{1, \cdots, n\}$ be an invertible map, initialize $\bar{E}$ to be the set $E$
  for $i = n, \cdots, 2$ do
    if $v_j$, $v_k$ are neighbours of $\pi^{-1}(i)$ in $G$ and $\pi(v_j)$, $\pi(v_k) < i$ then
      $\bar{E} \leftarrow E \cup \{(v_j, v_k)\}$
    endif
  endo
end add_edge
```

To see that *add_edge* does the job, consider any cycle O in G. Let v_i be the vertex of largest value π on O and let v_j, v_k denote the neighbours of v_k on O. Clearly $(v_j, v_k) \in \bar{E}$ and hence O will need at least 3 colors when $\bar{G}$ is colored.

It is clear that the initial ordering π will affect the resulting graph $\bar{G}$ and consequently the number of colors used. For example, if G is the wheel graph on 9 vertices shown in Fig. 7, and if the center vertex is ordered last, then $\bar{G}$ is a complete graph and requires 9 colors. On the other hand, if the center vertex is ordered first, and the outer vertices are ordered sequentially, then $\bar{G}$ is constructed from G by adding an edge between v_2 and v_9; $\bar{G}$ requires just 4 colors in this case.

A successful heuristic labelling rule, suggested by Powell and Toint, is the following. Assume that the vertices $\pi^{-1}(n), \cdots, \pi^{-1}(n-k)$ have been found. Choose as the vertex to be ordered $n-k-1$, the vertex of smallest degree in $G-\{\pi^{-1}(n), \cdots, \pi^{-1}(n-k)\}$. This algorithm is known as the smallest last ordering (*slo*)

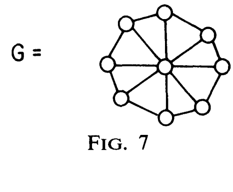


FIG. 7

and has a number of interesting properties. For further information consult Coleman and Moré [1984] and Matula and Beck [1983].

The algorithms of Coleman and Moré [1984] and Powell and Toint [1979] both implicitly perform *add_edge/slo* followed by a $\bar{G}$-coloring step. Here they differ: the latter authors apply colors in a greedy fashion by considering the nodes in the given order (i.e. $\pi^{-1}(1), \cdots, \pi^{-1}(n)$), Coleman and Moré apply a greedy algorithm over several different (cleverly chosen!) orderings. It has been proven that the coloring problem restricted to the class of graphs derived from the *add_edge/slo* completion process is NP-complete. However, an important question remains: Does an *optimal* coloring of such a completed graph always solve the cyclic coloring problem? If the answer is yes then one may conclude that it is not necessary to consider algorithms outside this framework. The answer is no.

To see that the cyclic coloring problem may not be solved by an *optimal* coloring of a completion produced by *add_edge*, consider Fig. 8.

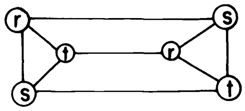


FIG. 8

The assignment shown is a valid cyclic 3-coloring; however, since every vertex is of degree 3 it follows that a coloring of a graph completed by algorithm *add_edge* will use at least 4 colors regardless of the ordering, π, of the nodes.

This example suggests that it may be worthwhile investigating heuristic algorithms for the partition/substitution problem, based on the cyclic coloring characterization perhaps, but not of the *add_edge* variety. At this point we do not know whether there is a practical gain to be made; the answer lies with further experimentation.

One final observation before discussing the solution process: A slight modification of the algorithm *add_edge* yields a procedure for the path coloring problem (symmetric direct method). In particular, change the conditional to read

"if v_j, v_k are neighbours of $\pi^{-1}(i)$ in G and $\pi(v_k) < i$ then"

and it follows that a color assignment of $\bar{G}$ is a path coloring of G. (To our knowledge this heuristic has not been suggested or experimented with previously.) If the conditional is further changed to read

"if v_j, v_k are neighbours of v_i in G then"

it follows that a coloring of $\bar{G}$ is a coloring of G^2 (and hence is also a path coloring of G).

An important computational concern is this: Given that ϕ: $V \rightarrow \{1, \cdots, p\}$ is a cyclic coloring, is it possible to compute the actual matrix elements in space proportional to $|E|$? It turns out that we can answer this question in the affirmative without imposing any additional structure on ϕ. In particular we do not assume that ϕ is necessarily consistent with the algorithm *add_edge/slo* (Coleman, Garbow, and Moré [1985] discuss, in detail, a FORTRAN 77 implementation of *add_edge/slo*, followed by a graph coloring step. Their substitution process operates in space $O(|E|)$; however it relies heavily on the regular matrix structure produced by *add_edge/slo*).

We will assume that H is stored as a sparse matrix and hence the space required is $O(|E|)$ where it is assumed that $|E| \geqq n$. We will not discuss time complexity here since it is very difficult to present a convincing argument without discussing detailed data structure and implementation requirements: we prefer to provide this analysis in a subsequent paper describing a specific implementation along with numerical results. Our purpose here is to support the claim that excessive space is not required. This discussion can be given at a fairly abstract level.

The first job is to determine $Hd_j = \nabla f(x+d_j) - \nabla f(x) \triangleq u_j$ and to save the significant information (nonzeros), for $j = 1, \ldots, p$, where p is the number of colors used by ϕ. The vector d_j must be consistent with the color j: if S_j is the set of columns (nodes) in the jth group (color) then $d_j = \sum_{i \in S_j} h_i e_i$, where h_i is the steplength associated with column i. Assume that we have the vector u_j on tap. If $(u_j)_i$ is a nonzero then this quantity is stored as follows:

```
for each $k$ such that $H_{ik}$ is a nonzero do
  if $\phi(v_k) = j$ then $H_{ik} \leftarrow (u_j)_i$ endif
endo
```

We note that it is not really necessary to replicate the information in H as we have done here; however, not doing so requires a more complicated indexing scheme than we wish to describe here. The key point here is that the vector pairs (d_j, u_j) are processed sequentially and so only $2n$ space is required.

When the process is complete, H is fully assigned but the numbers do not correspond to the actual Hessian quantities: we must now effect a substitution process.

For any pair of colors r, s, we can extract a bi-colored tree, $T_{r,s}$ from the representation of G and store $T_{r,s}$ as a tree structure in space $O(n)$. Algorithm *solve_tree* can now be used: we need only be more specific about step *solve* (x_i, y_i). The idea is simply to effect (3.1) with the knowledge that the difference results u_j are stored in H (as indicated above). In particular, *solve* (x_i, y_i) should read

$$H_{ij} \leftarrow \frac{H_{ij} - \sum_{k \in N(i)} H_{ik} \cdot h_k}{h_j},$$

$$H_{ji} \leftarrow H_{ij}$$

where $N(i)$ is the index set of neighbours of vertex x_i in $T_{r,s}(x_i)$ (i.e. all neighbours of x_i in $T_{r,s}$ except y_i). The reason this works is that when H_{ij} is solved, all other elements in row i (of columns in the same group as column j) have already been resolved.

It follows that the space required to resolve all unknowns is $O(|E|)$.

6. Concluding remarks. We have analyzed a class of methods for estimating sparse Hessian matrices, namely, substitution methods. In particular we have shown that there is an easy and elegant graph theoretic characterization of all substitution procedures based on a partition of columns of the symmetric matrix H. This characterization has allowed for a rich understanding of the combinatorial nature of the problem: we have analyzed the complexity of the partition problem, as well as suggested efficient procedures to effect the substitution process.

We have restricted our attention, in this paper, to substitution procedures based on a *partition* of columns. Indeed this is more restrictive than need be: Powell and Toint [1979], in their example (5.3), demonstrated that allowing the assignment of a column to several groups can reduce the number of required gradient evaluations. This example is particularly interesting because the solution procedure remains a "substitution process" requiring no matrix factorization. However, since the procedure allows a column to belong to several groups, it is not a method based on a partition of columns and does not belong in the class of substitution methods considered in this paper.

Indeed, a more general scheme than even this is possible provided matrix factorizations are acceptable. Newsam and Ramsdell [1983] have explored this general "elimination" option (Coleman [1984] summarizes this idea on page 49). While such methods may occasionally yield a reduction in the number of gradient evaluations, it is not clear that they provide a net benefit, in general, since they require the solution of n square dense (but relatively small) systems of equations to recover the true information.

Two other works should be mentioned. Thapa [1984] has also suggested a direct/partition method for estimating sparse Hessian matrices. Goldfarb and Toint [1984] have proposed specific (optimal) substitution procedures for specific common "mesh structures": such procedures are, of course, efficient algorithms for obtaining and using optimal cyclic colorings for particular regular structures.

We end with a comment on parallelism. There is a high degree of parallelism in the Hessian estimation problem. Specifically, each estimation Hd_j, $j=1,\cdots,p$ can be done independently, and thus in parallel. Since this work is sometimes the dominant expense in a numerical problem, exploiting this concurrency may be quite profitable. Note that the number of processors would usually be quite modest, even for large problems, since a cyclic coloring typically uses much fewer than n colors. Moreover, the substitution process also allows for parallel computation: each bi-colored tree can be processed entirely independently of the others.

Acknowledgment. We are grateful to Jorge Moré whose many suggestions have improved this paper. Specifically, he deserves credit for the elegant 6-vertex graph in § 5 that replaces our original graph of 25 vertices.

REFERENCES

T. F. COLEMAN [1984], *Large Sparse Numerical Optimization*, Lecture Notes in Computer Sciences 165, Springer-Verlag, New York.

T. F. COLEMAN AND J. 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.

T. F. COLEMAN, B. GARBOW AND J. J. MORÉ [1984], *Software for estimating sparse Jacobian matrices*, ACM Trans. Mathematical Software, 10, pp. 329-347.

——— [1985], *Software for estimating sparse Hessian matrices*, Technical Report 43, Argonne National Laboratory, Argonne, IL. Also available as TR 85-660, Dept. of Computer Science, Cornell Univ., Ithaca, NY.

A. R. CURTIS, M. J. D. POWELL AND J. K. REID [1974], *On the estimation of sparse Jacobian matrices*, J. Inst. of Math. Appl., 13, pp. 117-119.

M. R. GAREY AND D. S. JOHNSON [1979], *Computers and Intractability, A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco.

P. E. GILL, W. MURRAY, M. A. SAUNDERS AND M. WRIGHT [1983], *Computing forward-difference intervals for numerical optimization*, SIAM J. Sci. Statist. Comp., 4, pp. 310-321.

D. GOLDFARB AND PH. L. TOINT [1984], *Optimal estimation of Jacobian and Hessian matrices that arise in finite difference calculations*, Math. Comput., 43, pp. 69-88.

D. W. MATULA AND L. L. BECK [1981], *Smallest-last ordering and clustering and graph coloring algorithms*, J. Assoc. Comput. Mach., 30, pp. 417-427.

S. T. MCCORMICK [1983], *Optimal approximation of sparse Hessians and its equivalence to a graph coloring problem*, Math. Programming, 26, pp. 153-171.

G. N. NEWSAM AND J. D. RAMSDELL [1983], *Estimation of sparse Jacobian matrices*, this Journal, 4, pp. 404-418.

M. J. D. POWELL AND PH. L. TOINT [1979], *On the estimation of sparse Hessian matrices*, SIAM J. Numer. Anal., 16, pp. 1060-1074.

M. N. THAPA [1984], *Optimization of unconstrained functions with sparse Hessian matrices—Newton-type methods*, Math. Programming, 29, pp. 156-186.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

008

DIFFERENCE METHODS FOR THE NUMERICAL SOLUTION OF TIME-VARYING SINGULAR SYSTEMS OF DIFFERENTIAL EQUATIONS*

KENNETH D. CLARK†

Abstract. In this note, we introduce a class of difference methods for the numerical solution of differential equations of the form

$$A(t)x'+B(t)x(t)=f(t)$$

where A, B, and f are assumed sufficiently smooth in t in the interval $I=[0, T]$ and $A(t)$ is identically singular on I. These methods are straightforward extensions of the well-known Gear's backward difference methods (BDF's) and correspond to BDF's whenever A is constant. It is shown that the modified methods (MBDF's) work whenever the system can be transformed to a constant coefficient problem by a change of variable $x=Ly$, and also whenever a related system can be transformed into a certain canonical form.

We also investigate the relationship between the convergence of BDF's and the continuous regularization of the system by its pencil perturbation. In particular, we show the existence of examples where the BDF's converge but the pencil perturbation is not a continuous regularization.

Key words. backwards difference formula, singular linear system, pencil perturbation

AMS(MOS) subject classifications. 34A08, 65L05

1. Introduction. The numerical solution of singular systems of differential equations of the form

$$A(t)x'(t)+B(t)x(t)=f(t) \tag{1.1}$$

(also commonly known as differential-algebraic equations (DAE's), implicit ODE's, semistate equations, and descriptor systems) has been a topic of considerable interest in recent years [13], [14], [32], [34]. Gear, in his original paper [21], introduced the use of implicit, multistep methods or backward difference formulas (BDF) on systems in *semi-explicit* form [3],

$$x'(t)=f(x,y,t), \qquad 0=g(x,y,t) \tag{1.2}$$

under the assumption that $g_y(x, y, t)$ is nonsingular. Equations of the form (1.1), (1.2) occur in optimal control problems [3], [10], in electric circuit problems [10], [13], [30], and in fluid dynamics [1], [26], [31]. More recently, circuit problems involving operational amplifiers have been shown to lead to equations similar to (1.2) with $g_y(x, y, t)$ singular [14]. These are the so-called *higher index* problems. Other examples of applications of higher index problems may be found in [7], [13], [26].

While the analytic and numerical theory of (1.1) (see [2], [3], [5], [13], [14], [33], [34]) has undergone substantial development, most of the research has concentrated on the constant coefficient problems [13], [34], [35], and problems of the form (1.2) under special assumptions (e.g., g_y nonsingular, g independent of y, $g_x * f_y$ nonsingular) [3], [21], [27]-[29]. In fact, conditions for the existence and uniqueness of solutions to (1.1) have not yet been established in the general case.

* Received by the editors September 20, 1984, and in revised form April 17, 1985. This research supported in part by the National Science Foundation under grant DMS-8318026, and the Air Force Office of Scientific Research under grant AFOSR-84-0240.

† Departments of Computer Science and Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8206. The author is currently a Ph.D. candidate in Applied Mathematics at North Carolina State University under the direction of S. L. Campbell.

In the linear, constant coefficient case, the implicit, fixed time-step, k-step methods introduced by Gear have been shown to be convergent and stable, independent of the stepsize, for $k<7$ [25], [14], [34]. These methods also work for several time-varying linear and nonlinear problems (see [3], [8]), but as is well known, they do not possess the properties of convergence and stability for the general problems (1.1) or (1.2).

In this paper, we introduce difference methods for numerically solving (1.1), and show that they work on a reasonably large class of time-varying problems (including many examples frequently appearing in the literature) for which the usual Gear's methods need not work. The methods we derive are a natural extension of Gear's methods and reduce to these methods when A is constant.

2. Notation and terminology. Following [8], we say that (1.1) is *analytically solvable* on an interval $I=[0, T]$ if for sufficiently smooth $f(t)$ solutions to (1.1) exist and are uniquely determined by their values at any $t_0 \in I$. A vector x^0 is a *consistent initial value* for (1.1) if there exists a functional solution $x(t)$ of (1.1) such that $x(t_0)=x^0$. If there is a scalar function $d(t)$ such that $(d(t)A(t)+B(t))^{-1}$ exists on I, then the matrix pair (or pencil) (A, B) is *regular*. Otherwise, the pencil is *singular*. For constant coefficient problems (1.1), regularity of the pencil (A, B) is equivalent to analytic solvability, whereas for time-varying problems this is not the case [3], [14]. When $(A(t), B(t))$ is regular the *local index* of (1.1), $\operatorname{ind}(A(t), B(t))$, is defined to be the index of $(dA+B)^{-1}A$. See [13] or [15] for more details on the index of a matrix. The index 0 problem corresponds to $A(t)$ nonsingular while (1.2) is index one if and only if g_y is nonsingular. We are concerned with higher index problems ($\operatorname{ind}(A, B)>1$) since these are the problems on which the implicit BDF's are known to sometimes fail.

The equation

$$F(x_\varepsilon', x_\varepsilon, \varepsilon, t)=0, \qquad x_\varepsilon(0)=x^0 \tag{2.1}$$

is a *regularization* (or regularizing perturbation) of (1.1) if $F_{x_\varepsilon'}$ is nonsingular for $\varepsilon>0$, and $F(x', x, 0, t)=0$, $x(0)=x^0$ is equivalent to (1.1). If for every initial condition x^0, $\lim_{\varepsilon\to 0} x_\varepsilon = x$, where x_ε is a solution of (2.1) such that $x_\varepsilon(0)=x^0$, and x is a solution of (1.1), the regularization will be called *continuous* [11]. If the convergence is uniform (distributional, pointwise), the regularization will be called *u-continuous* (*d-*, *p-*continuous).

The concept of regularization is a generalization of the familiar singular perturbation for the constrained system (1.2):

$$\begin{aligned} x_\varepsilon' &= \tilde f(x_\varepsilon, y_\varepsilon, t, \varepsilon), \\ \varepsilon y_\varepsilon' &= \tilde g(x_\varepsilon, y_\varepsilon, t, \varepsilon) \end{aligned} \tag{2.2}$$

where $\tilde f|_{\varepsilon=0}=f$ and $\tilde g|_{\varepsilon=0}=g$.

A matrix C is *semistable* if and only if (i) $\operatorname{ind}(C)\leqq 1$, and (ii) $\lambda\in\sigma(C)$ and $\lambda\neq 0$ implies $\operatorname{Re}\{\lambda\}<0$. For vectors x, y define $(x, y)^t=(x^T, y^T)^T$, where $(\cdot)^T$ denotes the transpose.

3. Derivation of the methods. The modified difference formulas (MBDF) we present can be derived from a difference operator point of view, following the direction of [34], or more simply by imbedding (1.1) into a system of the form

$$\begin{aligned} z'(t) &= [A'(t)-B(t)]x(t)+f(t), \\ 0 &= z(t)-A(t)x(t) \end{aligned} \tag{3.1}$$

or equivalently $Pw'+Q(t)w=g$ where $w(t)=(z(t),x(t))^t$, $g(t)=(f(t),0)^t$, and

$$P=\begin{pmatrix} I_n & 0 \\ 0 & 0 \end{pmatrix}, \qquad Q(t)=\begin{pmatrix} 0 & B-A' \\ I_n & -A \end{pmatrix}. \tag{3.2}$$

(Note that solvability of (3.1) and (3.2) is slightly different since g is no longer arbitrary.) We then apply the Gear's formulas to (3.1) to obtain the modified methods. For example, the implicit 1-step method (Euler's) applied to (3.1) leads to the modified method

$$[A_{k+1}+h(B_{k+1}-A'_{k+1})]x_{k+1}=A_kx_k+hf_{k+1}. \tag{3.3}$$

Of course, (3.3) is a valid method only if $sA+(B-A')$ is invertible for some s. But this is equivalent to the regularity of (P, Q), as is easily shown by examination of the Schur complement of $sP+Q$. In [13], an example is given which establishes the independence of the regularity of (A, B) and (P, Q). However,

PROPOSITION 3.1. *The system* (1.1) *is analytically solvable if and only if* (3.1) *is analytically solvable.*

Proof. Suppose (3.1) is analytically solvable. Without loss of generality assume $t_0=0$. If $w^0=(z(0),x(0))^t=(z^0,x^0)^t$ is a consistent initial value for (3.1) then $z^0=A(0)x^0$. Let $w(t)=(z(t),x(t))^t$ be the unique solution of (3.1) such that $w(0)=w^0$. Differentiating $z=Ax$ and subtracting from equation one in (3.1) shows that $x(t)$ is a solution of (1.1) such that $x(0)=x^0$. This solution is unique since if $\mathbf{x}(t)$ is another such solution of (1.1), then $\mathbf{w}(t)=(A\mathbf{x},\mathbf{x})^t$ is a second solution of (3.1) satisfying $\mathbf{w}(0)=w^0$. This contradicts the solvability of (3.1). Hence, (1.1) is analytically solvable.

Now suppose (1.1) is analytically solvable, $x(t)$ is a solution, and $x(0)=x^0$. Clearly $w(t)=(A(t)x(t),x(t))^t$ solves (3.1) and $w(0)=(A(0)x^0,x^0)^t$. Suppose there exist $\mathbf{w}(t)=(z(t),x(t))^t$ solving (3.1) with initial value w^0. If $\mathbf{x}(t)\neq x(t)$, then (1.1) is not solvable. But $\mathbf{x}(t)=x(t)$ implies $\mathbf{z}=A\mathbf{x}=Ax=z(t)$, thus implying solvability of (3.1). QED.

Since $A(t)$ is singular the system (3.1) is higher index. In [22], [33], the *global index* of a system (1.1) is defined in terms of its reduction to certain canonical forms. An algorithm is given in [33] for determining the global index of (1.1), and for $A(t)$, $B(t)$ analytic the algorithm terminates in m steps if and only if the global index is m. Using this algorithm we can easily prove

PROPOSITION 3.2. *If the global index* (1.1) *is* m, *then global index* (3.1) *is* $\dot{m}+1$.

Proof. Differentiate the second equation in (3.1) to obtain the system

$$\begin{pmatrix} I_n & 0 \\ I_n & -A \end{pmatrix}\begin{pmatrix} z' \\ x' \end{pmatrix}+\begin{pmatrix} 0 & (B-A') \\ 0 & -A' \end{pmatrix}\begin{pmatrix} z \\ x \end{pmatrix}=\begin{pmatrix} f \\ 0 \end{pmatrix} \tag{3.4}$$

completing the first step. Now perform elementary row operations on (3.4) to obtain

$$\begin{pmatrix} I_n & 0 \\ 0 & \begin{bmatrix} \tilde{A} \\ 0 \end{bmatrix} \end{pmatrix}\begin{pmatrix} z' \\ x' \end{pmatrix}+\begin{pmatrix} 0 & (B-A') \\ 0 & \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \end{pmatrix}\begin{pmatrix} z \\ x \end{pmatrix}=\begin{pmatrix} f \\ \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \end{pmatrix} \tag{3.5}$$

where $\left[\begin{smallmatrix} \tilde{A} \\ 0 \end{smallmatrix}\right]$ is the result of zeroing a maximum number of rows in A, and permuting the zero rows to the bottom. Differentiation of $B_2x=f_2$ completes the second step of the algorithm on (3.1) and the first step on (1.1). Since global index (1.1) is m, this procedure terminates in $m-1$ more steps with a nonsingular coefficient leading $({}^{\backprime}z', x')^t$. QED.

To obtain the general form of the modified methods, we apply the Gear's multistep formulas to system (3.1). From [34, p. 24], the k-step implicit method with constant stepsize h is

$$[P+hb_0Q_n]w_n=\sum_{i=1}^{k} a_iPw_{n-i}+hb_0g_n \tag{3.6}$$

where $b_0=[\sum_{i=1}^{k}(1/i)]^{-1}$ and $a_i=(-1)^{i+1}b_0\sum_{j=i}^{k}\{(1/j)\binom{j}{i}\}$ and $\binom{j}{i}$ denotes the binomial coefficient "j lower i". This immediately leads to the modified k-step formula

$$[A_n+hb_0(B_n-A'_n)]x_n=\sum_{i=1}^{k} a_iA_{n-i}x_{n-i}+hb_0f_n. \tag{3.7}$$

Note that when A is constant, (3.7) is the usual k-step implicit method, so in this sense the modified methods are a natural extension of these methods. Also, if A is nonsingular, then (3.1) is index one so that the methods work in this case as well.

Example 1. Consider the system $Ax'+Bx=f$ where

$$A=\begin{pmatrix}0 & 0\\ 1 & vt\end{pmatrix},\quad B=\begin{pmatrix}1 & vt\\ 0 & 1+v\end{pmatrix},\quad f=(g,0)^T.$$

This example has been referred to often in the literature (see [3], [13], [22]). If g is differentiable, the exact solution to this system is

$$x_1(t)=g(t)+vtg'(t),\qquad x_2(t)=-g'(t).$$

If $v\neq -1$ we may apply an implicit Euler's scheme to get the system

$$x_{1,n}=g_n-vt_nx_{2,n},\qquad x_{2,n}=(v/v+1)x_{2,n-1}-[(g_n-g_{n-1})/h(1+v)]$$

which exhibits stability problems if $v<-\frac{1}{2}$. However, applying the modified one-step method yields the system of equations

$$x_{1,n}=g_n-vt_nx_{2,n},\qquad x_{2,n}=-(g_n-g_{n-1})/h$$

which, ignoring roundoff errors in numerically differentiating g, converges to the exact solution and is stable independent of v.

Note that for this problem, the change of variable $x=Ly$ with $L=\begin{pmatrix}1 & vt\\ 0 & 1\end{pmatrix}$ transforms the original system into a constant coefficient nilpotent, index two problem $Ny'+y=f$ with $N=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}$.

Many examples used in the literature to illustrate the poor behavior of Gear's methods on the general time-varying problem are obtained from constant coefficient problems by time-varying coordinate changes. Thus it seems reasonable to determine, if possible, a class of methods which can be used on these problems.

We say (1.1) is *transformable to constant coefficients* if there exists a nonsingular, differentiable $L(t)$ such that the substitution $x=Ly$ transforms (1.1) to a constant coefficient solvable system. Such transformable systems have been characterized in [13, p. 140] by the following theorem.

THEOREM 3.3. *The system* (1.1) *is transformable to constant coefficients if and only if* (i) $sA+B-A'$ *is invertible on* I *for some* s, *and* (ii) $A(sA+B-A')^{-1}$ *is constant on* I. *If* (i), (ii) *hold, we may take* $L=(sA+B-A')^{-1}$ *to obtain the system* $Cy'+(I-sC)y=f$, *where* $C=AL$. *Furthermore, for each* v *such that* $vA+B-A'$ *is invertible, the quantities* $C_v=A(vA+B-A')^{-1}$, $C_v^DC_v$, *and* $C_v^D(I-vC_v)$ *are independent of* t (*here,* $(\cdot)^D$ *denotes the Drazin inverse* [15]).

Often, premultiplication by a smooth matrix function $G(t)$ may transform a system which is not transformable to constant coefficients into a system which can be transformed to constant coefficients (see example 6.2.3 [13, p. 142]).

Let s_0A+B-A' be invertible on I for $s_0 \in \mathbb{C}$. Then there is a neighborhood of $\{0\}\times I$ in $\mathbb{C}\times I$ where $R(s,t)=[(s_0-s)A(t)+B(t)-A'(t)]^{-1}$ is an analytic matrix-valued function of s. But then

$$\begin{aligned} R(s,t) &= [(s_0-s)A(t)+B(t)-A'(t)]^{-1} \\ &= [(I-sA(s_0A+B-A')^{-1})(s_0A+B-A')]^{-1} \qquad (3.8)\\ &= (s_0A+B-A')^{-1}[I-sC]^{-1} \end{aligned}$$

where C is independent of t by Theorem 3.3. If C is nilpotent, then $R(s,t)$ is analytic in s for every $(s,t)\in\mathbb{C}\times I$, and in particular for $s=s_0$. Hence, without loss of generality we may choose $s_0=0$. We have just proved

PROPOSITION 3.4. *Suppose* (1.1) *is transformable to a constant coefficient problem with C nilpotent. Then $B(t)-A'(t)$ is nonsingular on I. More generally, if s_0A+B-A' is invertible for some s_0, and $A(s_0A+B-A')^{-1}$ is nilpotent, then $B(t)-A'(t)$ is invertible.*

The converse to Proposition 3.4 is not true.

Example 2. Consider the nilpotent index two system $Nx'+x=f$ where

$$N(t)=2\begin{pmatrix} \cosh(t)\sinh(t) & \cosh^2(t) \\ -\sinh^2(t) & -\cosh(t)\sinh(t) \end{pmatrix}.$$

Then $I-N'$ is invertible on an interval about $t=0$ and

$$(I-N')^{-1}=-\tfrac{1}{3}\begin{pmatrix} 1+2\cosh(2t) & 2\sinh(2t) \\ -2\sinh(2t) & 1-2\cosh(2t) \end{pmatrix}.$$

However, the (1.1) entry of $N(I-N')^{-1}$ is $(-2/3)\cosh(t)\sinh(t)$ so that this problem is not transformable to constant coefficients.

Alternatively, if $B(t)-A'(t)$ is invertible on I and (1.1) is transformable to constant coefficients, then we may choose $s_0=0$ and hence the change of variable $x=Ly$, with $L(t)=(B(t)-A'(t))^{-1}$, transforms (1.1) into the constant coefficient problem $Cy'+y=f$. Here, C may not be nilpotent, but irrespective of this possibility we have

THEOREM 3.5. *Suppose* (1.1) *is transformable to constant coefficients and $B(t)-A'(t)$ is invertible on I. Then the methods* (3.7) *converge to a solution of* (1.1).

Before proving Theorem 3.5 we establish the following lemma, adopting the notation $L_n=L(t_n)=(B(t_n)-A'(t_n))^{-1}$, $t_n=nh$.

LEMMA 3.6. *If $\{y_n\}_0^N$ is a solution to the difference equation*

$$(C-hb_0I)y_n=\sum_{i=1}^{k} a_iCy_{n-i}+hb_0f_n \tag{3.9}$$

(*assume y_n is exact for $n=0,\cdots,k-1$*), *and $\{x_n\}_0^N$ is a solution of* (3.7) *such that $x_n=L_ny_n$ for $n=0,\cdots,k-1$, then $x_n=L_ny_n$ for $n=0,\cdots,N$.*

Proof. Note that $A=ALL^{-1}=CL^{-1}$. Thus (3.7) is equivalent to

$$[A_nL_n+hb_0I]L_n^{-1}x_n=\sum_{i=1}^{k} a_iCL_{n-i}^{-1}x_{n-i}+hb_0f_n \tag{3.10}$$

or

$$[C+hb_0I]L_n^{-1}x_n=\sum_{i=1}^{k} a_iCL_{n-i}^{-1}x_{n-i}+hb_0f_n. \tag{3.11}$$

Define $u_n=L_n^{-1}x_n$. Then $u_n=y_n$ for $n=0,\cdots,k-1$ and u_n satisfies (3.9) so that $u_n=y_n$ for each $n=0,\cdots,N$. The proposition follows immediately. QED.

Proof of Theorem 3.5. Since C is constant the methods (3.9) are convergent and stable independent of constant h for $k<7$. Also, since the transformation $x=Ly$ does not depend on h, the values $\{x_n\}$ must also converge. QED.

The choice $s_0=0$ merely simplifies the notation in lemma 3.6 and in fact the same relationship between (3.7) and the differences

$$[C+hb_0(I-s_0C)]y_n=\sum_{i=1}^{k} a_iCy_{n-i}+hb_0f_n \tag{3.12}$$

holds for any s_0 for which s_0A+B-A' is invertible. We state without proof

THEOREM 3.7. *If* (1.1) *is transformable to constant coefficients, then the methods* (3.7) *converge to a solution of* (1.1).

It is notable that many of the properties of Gear's methods for constant coefficient problems also hold for (3.7) for systems transformable to constant coefficients. For example, if m is the nilpotency of the constant coefficient problem and k is the order of the method, then after $(m-1)k+1$ steps, (3.12) tracks a solution corresponding to a consistent initial value [34, p. 24]; hence so do methods (3.7). Also, if $sA+B-A'$ is invertible for some s_0, then it is invertible for all but a finite number of values s_0. Thus there is some freedom in the selection of a well-condition transformation. When $B-A'$ is invertible, the choice $s_0=0$ is probably the best one.

The applicability of MBDF is not restricted to problems transformable to constant coefficients [9]. In [8], a canonical form, the *standard canonical form of size r* (SCF $-r$), is introduced, and it is shown that implicit BDF's work on systems which can be *safely transformed* into SCF $-r$; i.e., by constant coordinate changes $x=Qy$, or by premultiplication by smooth, nonsingular $P(t)$ (only the proof for implicit Euler's is given). Since MBDF on (1.1) is equivalent to BDF on (3.1), it suffices that (3.1) be safely transformable to SCF $-r$. It is easy to show this implies that the system $ALz'+(I-s_0AL)z=q_1$ be safely transformable to SCF $-r$ (L is the same as in Theorem 3.3).

However, there are examples where BDF's work and MBDF's cannot be used.

Example 3. Consider the system $Nx'+x=f$ where $N=\left(\begin{smallmatrix}0&1\\0&0\end{smallmatrix}\right)$. Premultiply this system by $G(t)=\left(\begin{smallmatrix}1&0\\t&1\end{smallmatrix}\right)$ to obtain the system

$$\begin{pmatrix}0&1\\0&t\end{pmatrix}x'+\begin{pmatrix}1&0\\t&1\end{pmatrix}x=\mathbf{f}. \tag{3.13}$$

Since BDF's work on constant coefficient problems, they also work on (3.13) (premultiplication by smooth invertible $G(t)$ preserves this property). But

$$A+h(B-A')=\begin{pmatrix}h&1\\ht&t\end{pmatrix}$$

is singular for every t, h; hence MBDF's are not applicable.

4. Numerical results. All computations were performed in APL in double precision ($E-16$) arithmetic on an IBM 370. The MBDF's and BDF's were compared on several examples, including examples nontransformable to constant coefficients, nonsingular, and index one problems. Some results of our numerical tests for the first order methods are tabulated below, and in order to make the presentation as compact as possible, all examples are integrated from $t=0$ to $t=1$ and three significant digits (rounded) of the l_1 norm of the error ($\max_i |e_i|$) is given for each method at $t=1$. "*****" indicates overflow.

Example 4. ([9]) $A(t)$, $B(t)$ as in example 2, $f(t)=P(t)(6t^2-t, t^3)^T$, where

$$P(t)=\begin{pmatrix} \cosh(t) & -\sinh(t) \\ -\sinh(t) & \cosh(t)\end{pmatrix},$$

the exact solution is $x(t)=(t\cosh(t)-t^3\sinh(t), t^3\cosh(t)-t\sinh(t))^T$, $x_0=(0,0)^T$.

h	MBDF	BDF
.05	3.98 E−01	3.27 E28
.01	4.07 E−02	******

As has already been indicated, this is an index two problem which is not transformable to constant coefficients. However, the slow growth of the error suggests that the applicability of MBDF may not be limited to transformable problems. In fact, this behavior of the error is expected due to the exponential nature of the solution.

Example 5.

$$A(t)=\begin{pmatrix} -2t & -1 \\ 4t^2 & 2t\end{pmatrix}, \quad B(t)=I, \quad x(t)=(\exp\{t\}, 1+t^2), \quad x_0=(1,1)^T,$$

$$f(t)=([1-2t]\exp\{t\}-2t-1, 4t^2\exp\{t\}+5t^2+2t+1)^T.$$

This local index two example is interesting for several reasons. First note that for any s, $\det(sN+I-N')=-3$. Hence $I-sC$ is invertible for every $s\in\mathbb{C}$, implying C is nilpotent. Taking $s=0$, it is easy to show that $N(I-N')^{-1}=-N$. Thus the system is not transformable to constant coefficients and in effect, using MBDF on the system is equivalent to using BDF on another nilpotent index two time-varying problem so that poor numerical behavior is expected.

h	MBDF	BDF
.05	5.44 E06	2.31 E07

This example shows that MBDF is not a general method for index two problems.

In the nonsingular and index one examples run, while both methods converged, the BDF methods were more accurate. This experience, combined with the additional cost of computing $A'(t)$, suggests that the MBDF methods offer an alternative approach to numerically solving (1.1) only for the higher index problems.

5. Comments on the practical implementation of MBDF. In practice, the symbolic differentiation of a matrix is a difficult and time-consuming process, especially if the matrix is large and its entries are complicated functions. The utility of MBDF as a general time-varying ODE solver (singular or nonsingular), if any, probably involves its implementation in conjunction with numerical difference schemes for $A'(t)$. By doing so, a new class of "hybrid" methods can be derived. A first order difference on $A(t)$ with the k-step MBDF leads to the mixed time-step scheme

$$[(1-b_0)A_n+b_0A_{n-1}+hb_0B_n]x_n=\sum_{i=1}^{k} a_iA_{n-i}x_{n-i}+hb_0f_n. \tag{5.1}$$

Since $b_0\in(0,1]$, (5.1) is BDF with A_n replaced by a convex combination of A_n and A_{n-1}, and with A_n replaced by A_{n-i} in the summation. In the case $k=1$ (i.e., $b_0=1$)

the method is

$$[A_{n-1}+hB_n]x_n = A_{n-1}x_{n-1}+hf_n \tag{5.2}$$

which is almost an explicit Euler's. The more general hybrid method is obtained by using a j-step difference on $A(t)$, yielding

$$\left[(1-b_0a_0)A_n - \left\{b_0 \sum_{p=1}^{j} a_pA_{n-p}\right\} + hb_0B_n\right]x_n = \sum_{i=1}^{k} a_iA_{n-i}x_{n-i} + hb_0f_n. \tag{5.3}$$

Note that (5.1) suggests we might consider b_0 as a free parameter in $(0, 1]$ and use this value to control local errors. The ensuing method would be neither variable-step nor variable-order, although for some singular problems (e.g., index one, and some index two problems) it might be possible to integrate b_0 into a variable-step variable-order scheme without creating instability. As of this time, we have not studied the feasibility of this type of strategy, nor have we pursued the use of methods (5.3). It is likely that in order to obtain a specified global accuracy we must require $j \geqq k$ in (5.3).

There are, on the other hand, instances where MBDF may have advantages over alternative methods. In [4], [6] a class of methods known as ith order, jth block ((i, j)) methods, is derived from series expansions. These methods are shown to work in all cases where (1.1) is known to be analytically solvable, and they do not suffer from ill-conditioning problems for small step sizes, as do BDF's. However, the (i, j) methods require the solution of an $(nj)\times(nj)$ singular linear system of equations at each time step (where j increases as the index of the problem increases), whereas MBDF only requires solving an $n\times n$ system at each step. Also, the coefficient matrix in the (i, j) methods involves $j-1$ derivatives of A and $j-1$ derivatives of B, where MBDF utilizes only $A'(t)$. The involvement of A' is not surprising since from [5], for some higher index problems, the solutions of (1.1) may depend directly on derivatives of the coefficient matrices. Finally, the (i, j) methods cannot indicate distributional behavior arising from inconsistent initial values, whereas MBDF is able to do so for problems transformable to constant coefficients.

Given (1.1), the two types of transformations one can perform on this system are coordinate changes and premultiplication by $Q(t)$. From the point of view of preserving numerical behavior, MBDF's share a triangular relationship with BDF's and the (i, j) methods. Under smooth coordinate changes and premultiplication by a nonsingular matrix, the (i, j) methods will work on the transformed system. From example 3, we have seen that premultiplication by a time-varying matrix can cause MBDF's to fail, but smooth coordinate changes preserve convergence and stability. For BDF's, the situation is reversed.

6. Pencil perturbations. An alternative approach to studying (1.1), (1.2) is through singular perturbations or regularizations. Here, the given system is perturbed to a nonsingular system by the introduction of parasitics. The solutions of the perturbed system are then analyzed as the perturbation parameters approach 0. Or, systems exhibiting fast dynamics or parasitic behavior are approximated by the corresponding *reduced order model*, i.e., the system obtained by setting the parasitic parameters to 0. Often, this reduced order model is a singular system (1.1) or (1.2).

For the most part, the numerical and perturbation theories for (1.1) have evolved independently. When A, B are constant, Cobb [18] has shown that the *pencil perturbation* for (1.1) (assuming regularity of the pencil (A, B))

$$(A+\varepsilon B)x'_\varepsilon + Bx_\varepsilon = f(t) \tag{6.1}$$

is a continuous regularization of (1.1). For the SCF$-r$ mentioned in § 3, it is proved in [8] that the pencil perturbation is a continuous regularization. Index one problems fall into this category.

There are problems for which the behavior of BDF's is known but for which no pencil perturbation-type results have been obtained. See, for example, Brenan's thesis on index 2 and 3 systems in the form (1.2) [3], or the work of R. März in [27]-[29] on nonlinear index 1 systems in fully-implicit form $F(x', x, t)=0$.

It has been observed by Lötstedt and Petzold [26] that the problems attendant to the implementation of BDF's on singular systems are mainly due to ill-conditioning of the matrix $A+hb_0B$ as $h\to 0^+$. Thus it seemed reasonable that the conditioning of $A+hb_0B$ and $A+\varepsilon B$ as h, $\varepsilon\to 0^+$ were intimately related. In fact, at one point the author believed that the convergence of BDF's was either a necessary or sufficient (or perhaps both) condition for the pencil perturbation to be a continuous regularization. However, as we show momentarily, there are problems for which BDF's are convergent and yet the pencil perturbation is not a continuous regularization.

The pencil perturbation for (3.1) is

$$\begin{pmatrix} I & \varepsilon(B-A') \\ \varepsilon I & -\varepsilon A \end{pmatrix}\begin{pmatrix} z'_\varepsilon \\ x'_\varepsilon \end{pmatrix}+\begin{pmatrix} 0 & (B-A') \\ I & -A \end{pmatrix}\begin{pmatrix} z_\varepsilon \\ x_\varepsilon \end{pmatrix}=\begin{pmatrix} f \\ 0 \end{pmatrix}. \tag{6.2}$$

Suppose (1.1) is transformable to constant coefficients. Then, from § 3, BDF's converge for (3.1). Suppose the transformation $x(t)=L(t)y(t)$, where $L(t)=(s_0A+B-A')^{-1}$, yields the constant coefficient system

$$Cy'+(I-s_0C)y=f \tag{6.3}$$

where $C=AL$, and assume $f(t)\equiv 0$ on I. By elementary row operations we may reduce (6.2) to the equivalent system

$$\begin{pmatrix} z'_\varepsilon \\ x'_\varepsilon \end{pmatrix}+\begin{pmatrix} (B-A')R_\varepsilon & 0 \\ -R_\varepsilon/\varepsilon & I/\varepsilon \end{pmatrix}\begin{pmatrix} z_\varepsilon \\ x_\varepsilon \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{6.4}$$

where $R_\varepsilon=[A+\varepsilon(B-A')]^{-1}=[(A/\varepsilon)+B-A']^{-1}/\varepsilon$. Hence, $AR_\varepsilon=C/\varepsilon$ independent of t by assumption.

From [13], the exact solution of (6.3) is

$$y(t)=\exp\{-C^D(I-s_0C)t\}CC^Dq \tag{6.5}$$

where q is an arbitrary vector. Hence, the exact solution of (3.1) is

$$x(t)=L(t)y(t)=(s_0A+B-A')^{-1}\exp\{-C^D(I-s_0C)t\}CC^Dq, \tag{6.6a}$$

$$\begin{aligned} z(t)&=A(t)x(t)=C\exp\{-C^D(I-s_0C)t\}CC^Dq \\ &=\exp\{-C^D(I-s_0C)t\}C^2C^Dq \end{aligned} \tag{6.6b}$$

since C commutes with C^D. In general $z(t)$ is not zero.

Note that z_ε in (6.4) is independent of x_ε. The equation for z_ε is

$$z'_\varepsilon+(B-A')R_\varepsilon z_\varepsilon=0$$

or, using properties of C,

$$z'_\varepsilon=\{(-I/\varepsilon)+(C/\varepsilon^2)\}z_\varepsilon \tag{6.7}$$

which has the solution

$$z_\varepsilon(t)=\exp\{(-I/\varepsilon)+(C/\varepsilon^2)]t\}z^0_\varepsilon. \tag{6.8}$$

Campbell shows in [12] that under certain hypotheses, matrix exponentials of the type in (6.8) have limits as $\varepsilon \to 0^+$. For our purposes, consider the following modified version of Theorem 1 in that paper, adopting the notation $[X; Y] = (I - Y^D Y) X (I - Y^D Y)$ for $n \times n$ matrices X, Y.

THEOREM 6.1. *If* $\operatorname{ind}(C) = 1$ *and C is semistable, then z_ε converges pointwise to* $z_0(t) = 0$ *for* $t > 0$.

Proof. From [12], under the assumption C is semistable, $\operatorname{ind}(C) = 1$, z_ε converges pointwise if and only if $[-I; C]$ is semistable. But $[-I; C] = (I - C^D C)(-I)(I - C^D C) = -(I - C^D C)$, since $C^D C$ (and hence $(I - C^D C)$) is a projector, so that $[-I; C]$ is index 1 with spectrum $\{0, -1\}$. Hence $[-I; C]$ is semistable and z_ε has a pointwise limit for $t > 0$. The limit of the exponential is calculated by using (7), (8) from [12] to yield

$$\begin{aligned} &\exp\{([[0; C]; [-I; C]] - [[C^D; C]; [-I; C]])t\}(I - [-I; C]^D[-I; C])(I - C^D C) \\ (6.9) \qquad &= (I - \{(I - C^D C)(-I)(I - C^D C)\}^D[-I; C])(I - C^D C) \\ &= \{I - (I - C^D C)^D (I - C^D C)\}(I - C^D C) \end{aligned}$$

since $[0; Y] = 0$ for any Y. But for a projector P, $P^2 P^D = P$ implies $PP^D = P^D P = P$. Hence (6.9) is

$$\{I - (I - C^D C)\}(I - C^D C) = C^D C (I - C^D C) = 0. \qquad \text{Q.E.D.}$$

Example 6. Let

$$A(t) = \begin{pmatrix} 1 & t \\ 0 & 0 \end{pmatrix}, \quad B(t) = \begin{pmatrix} -1 & 1-t \\ 0 & 1 \end{pmatrix}.$$

The system $Ax + Bx = 0$ is transformable to the constant coefficient system $Cy' + y = 0$ where $C = \left(\begin{smallmatrix} -1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, by letting $L(t) = (B - A')^{-1} = \left(\begin{smallmatrix} -1 & -t \\ 0 & 1 \end{smallmatrix}\right)$ (here $s_0 = 0$). The solution is $y(t) = \exp\{-C^D t\} C C^D q = \exp\{t\}\left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right) q$. Thus $z(t) = \exp\{t\}\left(\begin{smallmatrix} -1 & 0 \\ 0 & 0 \end{smallmatrix}\right) q$.

Note that C in example 6 is semistable and $z(t) \neq 0$ unless $q = (0, a)^T$. Thus BDF's converge for the (3.1) version of example 6, but the solutions to (6.2) for this example do not converge to the correct limits, unless z_ε^0 is $O(\varepsilon)$ as $\varepsilon \to 0^+$. Hence, the convergence of BDF's is, in general, not a sufficient condition for (6.1) to be a continuous regularization of (1.1).

Acknowledgments. The author would like to thank Stephen L. Campbell for many helpful suggestions, and the referees for pointing out some errors in numerical examples.

REFERENCES

[1] R. AMIT, C. A. HALL AND T. A. PORSCHING, *An application of network theory to the solution of implicit Navier-Stokes difference equations*, J. Comp. Phys., 40 (1981), pp. 183-201.

[2] ROBERT K. BRAYTON, FRED G. GUSTAVSON AND GARY D. HACHTEL, *A new and efficient algorithm for solving differential-algebraic systems using implicit backward differences formulas*, Proc. IEEE, 60 (1972), pp. 98-108.

[3] KATHRYN E. BRENAN, *Stability and convergence of difference approximations for higher index differential-algebraic systems with applications in trajectory control*, Ph.D. dissertation, Univ. California, Los Angeles, 1983.

[4] S. L. CAMPBELL, *Explicit methods for solving singular differential equation systems*, preprint, Dept. Mathematics, North Carolina State Univ., Raleigh, 1984.

[5] ———, *Index two linear time-varying singular systems of differential equations*, this Journal, 4 (1983), pp. 237-243.

[6] S. L. CAMPBELL, *Non*-BDF *methods for the solution of linear time varying implicit differential equations*, Proc. 1984 Amer. Cont. Conf., to appear.

[7] ———, *The numerical solution of differential algebraic equation systems*, preprint, Dept. Mathematics, North Carolina State Univ., Raleigh, 1984.

[8] ———, *One canonical form for higher-index linear time-varying singular system*, Circ. Sys. Sig. Proc., 3 (1983), pp. 311-326.

[9] ———, *personal communication.*

[10] ———, *A procedure for analyzing a class of nonlinear semistate equations that arise in circuit and control problems*, IEEE Trans. Circ. Sys., CAS-28 (1981), pp. 256-261.

[11] ———, *Regularizations of linear time-varying singular systems*, Automatica, 20 (1984), pp. 365-370.

[12] ———, *Singular perturbation of autonomous linear systems*, II, J. Differential Equations, 29 (1978), pp. 362-373.

[13] ———, *Singular Systems of Differential Equations*, Research Notes in Mathematics, Vol. 40, Pitman, Marshfield, MA, 1979.

[14] ———, *Singular Systems of Differential Equations* II, Research Notes in Mathematics, Vol. 61, Pitman, Marshfield, MA, 1982.

[15] S. L. CAMPBELL AND C. D. MEYER, *Generalized Inverses of Linear Transformations*, Surveys and References Works in Mathematics, Vol. 4, Pitman, Marshfield, MA, 1979.

[16] S. L. CAMPBELL, C. D. MEYER AND N. J. ROSE, *Applications of the Drazin inverse to linear systems of differential equations with singular constant coefficients*, SIAM J. Appl. Math., 31 (1976), pp. 411-425.

[17] S. L. CAMPBELL AND N. J. ROSE, *Singular perturbation of autonomous linear systems*, SIAM J. Math. Anal., 10 (1979), pp. 542-551.

[18] D. COBB, *On the solutions of linear systems of differential equations with singular coefficients*, J. Differential Equations, 46 (1982), pp. 310-323.

[19] B. FRANCIS, *Convergence in the boundary layer for singularly perturbed equations*, Automatica, 18 (1982), pp. 57-62.

[20] C. W. GEAR, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[21] ———, *Simultaneous numerical solution of differential-algebraic equations*, IEEE Trans. Circ. Theory, CT-18 (Jan. 1971), pp. 89-95.

[22] C. W. GEAR AND L. R. PETZOLD, *Differential/algebraic systems and matrix pencils*, in Matrix Pencils, Lecture Notes in Mathematics 973, Springer-Verlag, New York, 1983, pp. 75-89.

[23] C. W. GEAR AND K. W. TU, *The effect of variable mesh size on the stability of multistep methods*, SIAM J. Numer. Anal., 11 (1974), pp. 1025-1043.

[24] C. W. GEAR AND D. S. WATANABE, *Stability and convergence of variable order multistep methods*, SIAM J. Numer. Anal., 11 (1974), pp. 1044-1058.

[25] E. HAIRER AND G. WARNER, *On the instability of the* BDF formulas, SIAM J. Numer. Anal., 20 (1983), pp. 1206-1209.

[26] P. LOTSTEDT AND L. R. PETZOLD, *Numerical solution of nonlinear differential equations with algebraic constraints*, Sandia Report SAND83-8877, Albuquerque, NM.

[27] R. MARZ, *Multistep methods for initial value problems in implicit differential-algebraic equations*, preprint No. 22, Humboldt-Universität zu Berlin, Sektion Mathematik, 1981.

[28] ———, *On difference and shooting methods for boundary value problems in differential-algebraic equations*, preprint No. 24, Humboldt-Universität zu Berlin, Sektion Mathematik, 1982.

[29] ———, *On initial value problems in differential-algebraic equations and their numerical treatment*, preprint No. 44, Humboldt-Universität zu Berlin, Sektion Mathematik, 1982.

[30] R. W. NEWCOMB, *The semistate description of nonlinear time-variable circuits*, IEEE Trans. Circ. Sys., CAS-28 (1981), pp. 62-71.

[31] J. F. PAINTER, *Solving the Navier-Stokes equations with* LSODI *and the method of lines*, Report UCID-19262, 1981, Lawrence Livermore National Laboratory, Livermore, CA.

[32] L. PETZOLD, *Differential/algebraic equations are not* ODE'*s*, SIAM J. Sci. Stat. Comput., 3 (1982), pp. 367-384.

[33] L. R. PETZOLD AND C. W. GEAR, ODE *methods for the solution of differential/algebraic systems*, Sandia Report SAND82-8051, Oct. 1982.

[34] R. F. SINCOVEC, B. DEMBART, M. A. EPTON, A. M. ERISMAN, J. W. MANKE AND E. L. YIP, *Solvability of large scale descriptor systems*, Final Report, Boeing Computer Services Company, Tukwila, WA, June 1979.

[35] J. W. STARNER, *A numerical algorithm for the solution of implicit algebraic-differential systems of equations*, Tech. Report No. 318, Dept. Mathematics and Statistics, Univ. New Mexico, May 1976.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

009

A PACKING PROBLEM YOU CAN ALMOST SOLVE BY SITTING ON YOUR SUITCASE*

DORIT S. HOCHBAUM† AND DAVID B. SHMOYS‡

Abstract. In this paper, we present a novel approach for approximating solutions to the bin-packing and machine scheduling problems. In obtaining our results, we exploit a certain dual relationship that exists between these two problems.

We introduce the notion of a dual approximation algorithm, where for the bin-packing problem, the aim is to find approximate packings where at most the optimal number of bins are used, but the bins are allowed to be filled beyond their capacity. For this approach, the objective is to minimize the tardiness of the machine that finishes last. For bin-packing instances where the size of each piece is at least $(1/3-\varepsilon)$ times the capacity of the bin, we give an approximation algorithm A_ε that is guaranteed to produce a solution where no bin contains more than $(1+3\varepsilon/2)$ times the bin capacity. Thus we have a family of dual approximation algorithms, dependent on the problem instance, where the "closer" the instance is to belonging to a class that can be solved in polynomial-time, the better performance is guaranteed.

Using this result, we construct an approximation algorithm for the minimum makespan scheduling problem, that always finds a schedule where all jobs are completed by $\frac{5}{4}$ times the best completion time.

1. Introduction.

In this paper we introduce a novel approach for approximating solutions to the bin-packing and machine scheduling problems. In obtaining our results, we exploit a certain dual relationship that exists between these two problems. As the title of this paper suggests, in our approach we "squish" judiciously selected items so that the deformed items can be packed efficiently. Once packed, they are restored to their original sizes, but too late—the suitcase is closed.

More formally, in a bin-packing problem we are given a collection of items of prescribed sizes to be packed into the minimum number of bins of some fixed capacity. A problem closely related, in a sense to be elaborated on later, is the problem of minimizing the makespan of an m-machine scheduling problem. Here we are given a collection of n jobs of prescribed processing times to be scheduled on m identical machines so that all machines finish processing by the earliest possible time. Notice that in the context of scheduling, the bin-packing problem can be viewed as finding the minimum number of machines needed to process all jobs within a certain deadline. For the remainder of this paper we shall always refer to the bin-packing problem in this context. The minimum makespan problem shall be denoted $P//C_{\max}$ [GLLR].

Traditionally, an approximation algorithm for the bin-packing problem produces a solution that is feasible, i.e., no bin is packed beyond its capacity, but the number of bins might well exceed the optimal number of bins. In contrast to this, our approach produces solutions that never use more than the optimal number of bins at the cost of packing some bins slightly beyond their capacity. We measure the performance of such an approximate bin-packing algorithm by the fractional excess capacity used. We believe that this approach can and should be adapted to other combinatorial optimization problems. For an arbitrary combinatorial optimization problem, the approach involves finding solutions that could only be better than the optimum, while sacrificing

* Received by the editors June 11, 1984, and in revised form February 8, 1985. This title is paraphrased after a title of Martin Gardner's column [Ga] discussing a two-dimensional packing problem.

† University of California, Berkeley, California 94720. Department supported in part by the National Science Foundation under grant ECS-8204695.

‡ Division of Applied Sciences, Harvard University, Cambridge, Massachusetts 02138. The work of this author was supported in part by the National Science Foundation by a graduate fellowship and under grants MCS-8311422 and DCR-83-02385, and in part by DARPA order 4031, monitored by Naval Electronic System Command under contract N00039-C-0235.

a certain degree of infeasibility. The quality of such solutions is evaluated by the degree of infeasibility possible. To further illustrate this approach, consider the traveling salesperson problem. In this case we would seek a tour that visits as many cities as possible, while its cost is no more than the cost of the optimal—and feasible—tour. We will refer to such algorithms as *dual approximation algorithms.* The term dual is used to highlight the analogous situation in mathematical programming where feasibility vs. superoptimality corresponds to primal feasibility vs. dual feasibility.

We introduce a family of strongly NP-complete bin-packing problems $\{P_\varepsilon\}$ and algorithms $\{A_\varepsilon\}$, parameterized by a continuous variable $\varepsilon > 0$; the performance guarantee of algorithm A_ε for problem P_ε depends monotonically on ε. More specifically, P_ε denotes the bin-packing problem restricted to instances where the processing requirements are all at least $(\frac{1}{3}-\varepsilon)d$ where d denotes the given deadline; the algorithm A_ε produces a schedule where all jobs are completed within $(1+3\varepsilon/2)d$. Notice that a novel feature of this approach is that the performance guarantee can be viewed as dependent on the particular processing requirements of the instance.

This result can be distinguished from previous results, that are all of the following flavor. Graham, for instance, showed that the so-called LPT heuristic for the minimum makespan scheduling problem on m machines has a performance guarantee of $\frac{4}{3}-1/3m$ [Gr1]. Thus, this can be viewed as a family of problems $\{P_m\}$ with a family of algorithms $\{A_m\}$ with performance guarantee depending on the *discrete* parameter m. Our continuous parameterization is also somewhat more natural, as it is common to think of the number of machines as fixed and the instances as varying in processing requirements.

Finally, we will show how this family of dual approximation algorithms $\{A_\varepsilon\}$ for the bin-packing problem can be transformed into a traditional approximation algorithm for the machine scheduling problem. This algorithm will always schedule the jobs so that all processing is completed within $\frac{5}{4}$ of the optimal completion time. The best algorithm, in terms of worst case performance, known for this problem is called MULTIFIT, and is due to Coffman, Garey, and Johnson [CGJ]; Friesen [FS] has shown that this algorithm always produces a schedule where all jobs are finished within $\frac{6}{5}$ of the optimum. Although our algorithm has an inferior bound, the proof of the guarantee does not rely on an extremely intricate use of weighting functions, as does the proof by Friesen.

We also tested empirically the performance of our machine scheduling algorithm as compared to that of LPT and MULTIFIT.

2. A well-solvable class of bin-packing problems. In this section we show that there is a polynomial-time algorithm for the class of bin-packing instances I_0, where we restrict the processing times $p_i \geqq d/3$. If we make this restriction somewhat stronger, and require that $p_i > d/3$, then it is a routine exercise to show that, since at most two jobs can be processed by a machine before the deadline, the problem can be reduced to a matching problem [PS, p. 245]. This result can be obtained by using a more direct approach which in fact, can handle more general instances. A natural approach to solving this problem is as follows: schedule as many machines as possible with 3 jobs, and then pair up the remaining jobs in an optimal manner. This method does not work; this can be seen by considering the set of processing requirements, $\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{2}{3}, \frac{2}{3}, \frac{2}{3}\}$ with a deadline of 1. This method does work, however, if we first schedule optimally the jobs with $p_j \geqq d/2$. The intuition behind this, is that these are the most problematic of the jobs, since any pair of the others can be scheduled together. See Fig. 1.

```
procedure pack.onethird(I)
  begin
    {Stage 1: pack all items bigger than $d/2$}
    until all jobs are shorter than $d/2$ do
      begin
        let $i$ be any job with $p_i \geqq d/2$
        let $j$ be the largest job that can be scheduled with $i$ by time $d$
        if no such $j$ exists then pack $i$ by itself
                else pack $i$ and $j$ together
      end
    {Stage 2: pack jobs of size $d/3$}
    while there are three jobs $i, j, k$ of size $d/3$
      pack $i, j, k$ together
    {Stage 3: pack remaining jobs in pairs}
    for all unpacked jobs pair them up and pack 2 per bin
  end
```

FIG. 1

It is an easy exercise to show the following result, and this is left to the reader.

THEOREM 1. *For any instance in I_0, the algorithm pack.onethird gives an optimal bin-packing schedule.*

3. Restricted classes of hard bin-packing problems. In the previous section we showed that the bin-packing problem could be solved in polynomial time for instances where each processing time $p_j \geqq d/3$. In this section we show that it is very unlikely that this result can be extended any further, by proving that the bin-packing problem restricted to the class of instances, I_ε, where each processing time $p_j \geqq (\frac{1}{3} - \varepsilon)d$, is strongly NP-complete for any fixed ε, $0 < \varepsilon < \frac{1}{3}$.

THEOREM 2. *The bin-packing problem, even restricted to instances in I_ε is strongly* NP-*complete, for any fixed ε, $0 < \varepsilon < \frac{1}{3}$.*

Proof. To prove the result, we reduce from the 3-partition problem. This problem can be stated as follows:

3-PARTITION.

INSTANCE. A finite set A of $3m$ elements, a bound B and a "size" $s(a)$ for each $a \in A$, such that each $s(a)$ satisfies $B/4 < s(a) < B/2$ and such that $\sum_{a \in A} s(a) = mB$.

QUESTION. Can A be partitioned into m disjoint sets $S_1, S_2, \cdots, S_m$ such that for $1 \leqq i \leqq m$, $\sum_{a \in S_i} s(a) = B$?

It is important to note that the constraints on the item sizes imply that in a feasible 3-partition each S_i must contain exactly 3 elements of A.

The bin-packing instance that we construct will have $3m$ jobs, corresponding to the $3m$ elements of A. Let $p(a)$ denote the processing requirement of the job corresponding to element $a \in A$, and set

$$p(a) = \frac{1}{3} - \varepsilon + \frac{s(a)}{B} \cdot 3\varepsilon$$

and set the deadline $d = 1$. It is important to note that for all $a \in A$, $p(a) \geqq \frac{1}{3} - \varepsilon$.

We show that the bin-packing instance constructed has a feasible schedule with m machines if and only if there is a 3-partition of the original instance. First we prove that any 3-partition immediately yields a schedule for m machines. Suppose that $S_1, S_2, \cdots, S_m$ is the 3-partition. This implies that for each i, $\sum_{a \in S_i} s(a) = B$. But then

$$\sum_{a \in S_i} p(a) = \sum_{a \in S_i} \left(\frac{1}{3} - \varepsilon + \frac{s(a)}{B} \cdot 3\varepsilon\right) = 1 - 3\varepsilon + \frac{3\varepsilon}{B} \sum_{a \in S_i} s(a) = 1 - 3\varepsilon + \frac{3\varepsilon}{B}(B) = 1.$$

This implies that the set of jobs corresponding to S_i can be scheduled on one machine within the deadline, and thus we have a feasible schedule for m machines in total.

It remains to be shown that if there is a feasible schedule for the jobs on m machines, then there is a feasible 3-partition. Observe that

$$\sum_{a\in A} p(a) = m - 3m\varepsilon + \frac{3\varepsilon}{B}\sum_{a\in A} s(a) = m - 3m\varepsilon + \frac{3\varepsilon}{B}mB = m.$$

By the pigeonhole principle, since there are m machines each scheduled for at most one unit of time, we see that each machine must be scheduled to process jobs for exactly one unit of time. For machine i, let the set of jobs scheduled to be processed by this machine be S_i. We have shown that $\sum_{a\in S_i} p(a) = 1$. Notice that

$$1 = \sum_{a\in S_i} p(a) = 1 - 3\varepsilon + \frac{3\varepsilon}{B}\sum_{a\in S_i} s(a),$$

and as a result,

$$\sum_{a\in S_i} s(a) = B.$$

Since the 3-partition problem is strongly NP-complete, this proves that the bin-packing problem restricted to instances in I_ε is strongly NP-complete, for any fixed ε, $0 < \varepsilon < \frac{1}{3}$. □

4. A family of approximation algorithms for restricted classes of bin-packing problems. In the previous section we showed that the bin-packing problem, restricted to instances in I_ε, $\varepsilon > 0$ remains NP-complete (where for $\varepsilon = 0$ it is polynomial). As a result, it is natural to consider approximation algorithms for these problems. The traditional approach to approximation algorithms for bin-packing, or for any combinatorial optimization problem, is to find an algorithm that finds a feasible solution for the problem, with cost as close to the optimum as possible. For the bin-packing problem, this means finding a schedule for the jobs so that each job is completed by the deadline, but the number of machines may be greater than is necessary. An equally valid approach, which we believe to be new, is to find a schedule that uses at most the minimum number of machines, $OPT_{BP}(I)$, but relaxes the requirement that all jobs finish by the deadline. The performance of the algorithm is measured by the amount of time beyond the deadline that some machine must run.

In this section we present a family of approximation algorithms, $\{A_\varepsilon | 0 < \varepsilon < \frac{1}{12}\}$, such that A_ε, when run on an instance I from I_ε, produces a schedule that uses no more than the minimum number of machines, $OPT_{BP}(I)$. Furthermore, every job is finished by $(1+3\varepsilon/2)d$, where d is the deadline specified by the instance.

Let the instance $I \in I_\varepsilon$ have n jobs, indexed $1, \cdots, n$, where the processing requirement of job i is p_i, and let the deadline be d. We first show that some of the jobs can still be packed optimally.

LEMMA 1. *For an instance of the bin-packing problem in I_ε, $0 < \varepsilon < \frac{1}{12}$, let i be a job with processing time $p_i \geqq d/2$. Let j be the longest job such that $p_i + p_j \leqq d$. Then there exists an optimal schedule that has i and j scheduled together as the only jobs processed by some machine.*

Proof. This claim is proved using a standard "interchange" argument. Let i and j be two jobs as specified in the hypothesis of the lemma. Suppose that there does not exist an optimal schedule where i and j are processed by the same machine. Consider any optimal schedule, and in particular, consider the schedule of the machine, M_i, that processes job i. Since all jobs have a processing requirement greater than $d/4$, machine M_l must be scheduled to process both i and at most one other job. If M_l only processes i, then clearly j can be added to M_l's schedule, which is a contradiction. So we can assume that M_l also processes job k. If we interchange j and k so that j is processed by machine M_l and k is processed by the machine that j had previously been scheduled on, we get a feasible schedule, since $p_k \leqq p_j$ (because j is the longest job that could be scheduled with i), and so $p_i + p_j \leqq p_k + p_j \leqq d$. This contradicts the original assumption. □

Therefore, we can preprocess our instance of the bin-packing problem to schedule all jobs with processing requirement at least $d/2$. Now we restrict our attention to instances where all of the processing times p_i fall in the interval $(d/4, d/2)$. It is obvious that any two jobs can be feasibly scheduled together, and at most three jobs can be scheduled together.

We define two sets of jobs which are parameterized by a variable δ:

$$\text{Largest}\,(\delta) = \left\{ j \,\middle|\, \left(\frac{1}{3} - \delta\right) d \leqq p_j \leqq \left(\frac{1}{3} + 2\delta\right) d \right\},$$

$$\text{Middle}\,(\delta) = \left\{ j \,\middle|\, \left(\frac{1}{3} - \delta\right) d \leqq p_j \leqq \left(\frac{1}{3} + \frac{\delta}{2}\right) d \right\}.$$

We first state some very simple facts.

FACT 1. *If there is an optimal schedule where one machine is scheduled with three jobs, then there exists an optimal schedule where the job with the minimum processing time is scheduled on some machine with two other jobs.*

FACT 2. *Let job* min *be the job with minimum processing time and suppose that* $p_{\min} = (\frac{1}{3} - \varepsilon)d$. *If* min *is scheduled with jobs i and j, where* $p_i \geqq p_j$ *then* $i \in \text{Largest}\,(\varepsilon)$ *and* $j \in \text{Middle}\,(\varepsilon)$.

Proof. Since $p_{\min} = (\frac{1}{3} - \varepsilon)d$, the sum $p_i + p_j \leqq (\frac{2}{3} + \varepsilon)d$, and by the pigeonhole principle, p_j is at most $(\frac{2}{3} + \varepsilon)d/2 = (\frac{1}{3} + \varepsilon/2)d$. Furthermore, since $p_j \geqq (\frac{1}{3} - \varepsilon)d$, it follows that $p_i \leqq d - 2(\frac{1}{3} - \varepsilon)d = (\frac{1}{3} + 2\varepsilon)d$. □

FACT 3. *If a bin-packing instance $\tilde{I}$ can be obtained from an instance I, by simply decreasing some of the processing requirements,* $OPT_{BP}(\tilde{I}) \leqq OPT_{BP}(I)$.

Consider the algorithm in Fig. 2.

Notice that the procedure does not necessarily provide a feasible schedule, in that some machines may be scheduled beyond the deadline. However, it is easy to see that no machine is scheduled with more than $(\frac{1}{3} - \varepsilon)d + (\frac{1}{3} + 2\varepsilon)d + (\frac{1}{3} + \varepsilon/2)d = (1 + \frac{3}{2}\varepsilon)d$, where ε is defined so that the shortest processing time is $(\frac{1}{3} - \varepsilon)d$. We now prove that, in fact, the number of machines used by the procedure, which we shall denote $pack(I)$, is at most $OPT_{BP}(I)$.

LEMMA 2. $pack(I) \leqq OPT_{BP}(I)$.

Proof. By Lemma 1, we know that the first **while** loop of the procedure is scheduling optimally, so we need only consider the remainder of the algorithm, where all of the job sizes are in the interval $(d/4, d/2)$.

The proof is by induction on the number of jobs remaining, n. If $n \leqq 2$ the proof is trivial, since the algorithm clearly uses only one machine, which is the optimal schedule, if any jobs remain at all.

Now assume that the claim holds for all instances where the number of jobs is less than n. Consider two basic cases; either there is an optimal schedule where a machine gets three jobs, or there isn't. In the latter case it is clear that the procedure *pack*.$\frac{3}{2}\varepsilon$ can use no more machines than the optimal schedule. (Perhaps some machine

```
procedure pack. $3/2\varepsilon(I)$
  begin
    $J := \{1, 2, \cdots, n\}$
    until all jobs are shorter than $d/2$ do
      begin
        let $i$ be any job with $p_i \geqq d/2$
        let $j$ be the longest job that can be scheduled with $i$ by time $d$
        {if none exists, let $j = \emptyset$}
        schedule $i$ and $j$ together on one machine
        $J := J - \{i, j\}$
      end
    {all jobs have processing times in the interval $(d/4, d/2)$}
    repeat
      begin
        let min be the shortest job in $J$
        if $p_{\min} > d/3$ then failure
          else
            begin
              define $\varepsilon \geqq 0$ so that $p_{\min} = (\frac{1}{3} - \varepsilon)d$
              if there do not exist 2 jobs other than min, one in each of
                Largest $(\varepsilon)$ and Middle $(\varepsilon)$
                then failure
                else
                  begin
                    let $i$ be the job in Largest $(\varepsilon)$ $(\neq \min)$ with the largest processing time
                    let $j$ be the job in Middle $(\varepsilon)$ $(\neq \min, i)$ with the largest processing time
                    schedule min, $i$, and $j$ together
                    $J := J - \{\min, i, j\}$
                  end
            end
      end
    until failure
    {all of the jobs remaining must be processed 2 per machine}
    pack the remaining jobs 2 per machine in some arbitrary way (with
    perhaps one machine with only one job)
  end
```

FIG. 2

is scheduled with 3 jobs by our algorithm, but that can only reduce the number of machines used.) Therefore, assume that there is an optimal schedule P with some machine scheduled with three jobs. Furthermore, by Fact 1, we can assume that the shortest job, $p_{\min} = (\frac{1}{3} - \varepsilon)d$, $\varepsilon \geqq 0$, is scheduled on a machine with 2 other jobs, k and l. Assume, without loss of generality, that $p_k \geqq p_l$.

By Fact 2, $k \in \text{Largest}(\varepsilon)$ and $l \in \text{Middle}(\varepsilon)$, and therefore, the algorithm does not detect failure in the first iteration of the **repeat** loop. Suppose that the algorithm schedules min, i and j on one machine where $p_i \geqq p_j$. By the choice of i and j, we know that $p_i \geqq p_k$ and $p_j \geqq p_l$. Let $J_1 = J - \{\min, i, j\}$ and $J_2 = J - \{\min, k, l\}$. It is straightforward to see that the instance I_1 corresponding to J_1 can be obtained from the instance I_2 corresponding to J_2 by decreasing some of the processing requirements, and thus $OPT_{BP}(I_1) \leqq OPT_{BP}(I_2)$. Consider the bin-packing instance I_1; by the inductive hypothesis, $pack(I_1) \leqq OPT_{BP}(I_1)$. It is also easy to see that $pack(I) = pack(I_1) + 1$ and $OPT_{BP}(I) = OPT_{BP}(I_2) + 1$. Combining the inequalities and equalities, we get that $pack(I) \leqq OPT_{BP}(I)$. □

As a result of Lemma 2, and the observation about the total time that any machine can be scheduled for using *pack.* $\frac{3}{2}\varepsilon$, we have shown the following theorem.

THEOREM 3. *The procedure pack.*$\frac{3}{2}\varepsilon$ *delivers a schedule for any instance* $I \in I_\varepsilon$, $0 \leqq \varepsilon < \frac{1}{12}$, *that uses no more than the minimum number of machines needed to complete the jobs within d, and this schedule ensures that all jobs are completed within* $(1+\frac{3}{2}\varepsilon)d$.

The procedure *pack.*$\frac{3}{2}\varepsilon$ can be viewed in a slightly different way. The procedure given above can be used to produce a modified instance of the bin-packing problem. For any job i that is packed during the **repeat** loop, let $\tilde{p}_i = d/3$. Otherwise, let $\tilde{p}_i = p_i$. The processing times $\tilde{p}_i$ define a perturbed instance $\tilde{I}$. Note that the schedule produced by *pack.*$\frac{3}{2}\varepsilon$ is feasible for $\tilde{I}$. Furthermore, it is not hard to see that this schedule is in fact optimal for $\tilde{I}$ (which is an instance that "almost" lies in I_0.) Thus, procedure *pack.*$\frac{3}{2}\varepsilon$ might be presented as follows:

```
procedure pack.$\frac{3}{2}\varepsilon(I)$
  begin
    form instance $\tilde{I}$ from $I$
    pack $\tilde{I}$ optimally
    interpret the schedule for $\tilde{I}$ as a schedule for $I$
end
```

This corresponds precisely to our suitcase analogy.

5. An approximation algorithm for minimizing the makespan. The study of performance guarantees for approximation algorithms originated with Graham [Gr], [Gr1] and this initial work analyzed two algorithms for $P\|C_{\max}$. Both of these algorithms are list scheduling algorithms (LS), where the jobs are first arranged in some order in a list, and then scheduled by assigning the next job on the list whenever a machine is idle. Graham showed that if no restriction is placed on the list, then the completion time of the list scheduling algorithm is at most $(2-1/m)OPT_{MM}(I)$ where $OPT_{MM}(I)$ denotes the completion time of the optimal schedule, given a set of jobs I. Graham also showed that if the next job to be scheduled is always the one with the largest processing time, the so-called LPT rule, then the schedule produced has a completion time of at most $\frac{4}{3}OPT_{MM}(I)$.

The following fact, which is used in proving the above results, will be important for our purposes.

LEMMA 3. *For any list scheduling algorithm, if j is the last job to be completed in the schedule produced, then this completion time T is at most* $(1/m)\sum_{i=1}^n p_i + ((m-1)/m)p_j$.

Proof. Since j was scheduled at time $T-p_j$, *all* processors were scheduled up to that point, and thus

$$\sum_{i=1}^n p_i = p_j + \sum_{i\neq j} p_i \geqq p_j + m(T-p_j).$$

By solving for T, we get the desired result. □

It is useful to observe that we do not need that the schedule was produced by a list scheduling algorithm, only that no machine is idle earlier than the starting time of the job that finishes last. It is easy to see that $mOPT_{MM}(I) \geqq \sum_{i=1}^n p_i$, and T, as defined above, is at least $OPT_{MM}(I)$. Using these simple observations, we get the following two corollaries.

COROLLARY 1.

$$\frac{1}{m}\sum_{j=1}^n p_j \leqq OPT_{MM}(I) \leqq \frac{1}{m}\sum_{j=1}^n p_j + \max_{j=1,\cdots,n} p_j.$$

COROLLARY 2. *If* $p = \max_j p_j$, *then any schedule produced by a list scheduling algorithm finished within* $OPT_{MM}(I) + ((m-1)/m)p$.

Thus, if all of the job times are small, these algorithms do very well. Recall that the approximation algorithms presented in the previous section perform well if all of the job times are big. This suggests the two-phase approach given in Fig. 3. In this procedure, we use binary search to identify the smallest deadline d, such that all of the jobs of length greater than $d/4$ can be scheduled using *pack.*$\frac{3}{2}\varepsilon$ on $\leqq m$ machines. Since we might not need the precise value of the minimum, we settle for some k iterations of binary search.

```
procedure schedule $(I, k)$
  $LB := \max\left\{\frac{1}{m}\sum_{j=1}^{n} p_j, \max_{j=1,\cdots,n} p_j\right\}$
  $upper := LB + \max_{j=1,\cdots,n} p_j$
  $lower := LB$
  for $iter := 1$ to $k$ do
    begin
      $d := \frac{upper + lower}{2}$
      $J := \{j \mid p_j > d/4\}$
      call pack.$\frac{3}{2}\varepsilon(J)$
      if no more than $m$ machines used then
        begin
          $upper := d$
        end
      else $lower := d$
    end
      $d_0 := upper$
  {for the best value $d_0$ recompute the schedule}
  $J := \{j \mid p_j > d_0/4\}$
  call pack.$\frac{3}{2}\varepsilon(J)$
  complete partial schedule using list scheduling
```

FIG. 3

We next prove some easy facts about this algorithm.

FACT 4. *Throughout the execution of the algorithm* $OPT_{MM}(I) \geqq lower$.

Proof. By Corollary 1, the initial value of *lower* is a valid lower bound on the optimum. Suppose that *lower* is updated during the algorithm to some value d. Then the procedure *pack.*$\frac{3}{2}\varepsilon$ was unable to pack the set of jobs on m machines (within the relaxed deadline), and by Lemma 2, it follows that no schedule exists for the jobs in J that completes all of the jobs by time d. Since J is a subset of the complete set of jobs to be scheduled, it follows that $OPT_{MM}(I) \geqq d$. □

FACT 5. *If every job is completed in less than* $\frac{5}{4}d_0$ *in the schedule produced by the algorithm, then the finish time exceeds the optimal finish time by less than* $(\frac{1}{4} + 2^{-k+1})OPT_{MM}(I)$.

Proof. Note that after iteration i of the **for** loop, the difference between *upper* and *lower* is at most $2^{-i}LB$. The relationship of the relevant parameters is depicted in Fig. 4. By this observation and Fact 4, the difference between the scheduled completion time and the optimal completion time is $\leqq 2^{-k}LB + d_0/4$. By observing that $d_0 \leqq (1+2^{-k})lower \leqq (1+2^{-k})OPT_{MM}(I)$, we get the claimed result.

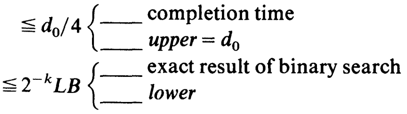


FIG. 4

FACT 6. *If the last job to finish is completed at some time* $T \geqq 5d_0/4$, *then the error is at most* $((m-1)/m)d_0/4 \leqq ((m-1)/4m + 2^{-k-2})OPT_{MM}(I)$.

Proof. By the arguments given in the previous section, we know that the subset of jobs J is scheduled within $5d_0/4$. Therefore, the job j that finishes last must be in $\{1, \cdots, n\} - J$, and no machine can be idle before the time that j is started. We know then that Lemma 3 is applicable, and the error is at most $((m-1)/m)p_j \leqq ((m-1)/m)d_0/4$. The final inequality is obtained by substituting $d_0 \leqq (1+2^{-k})OPT_{MM}(I)$. □

By combining Facts 5 and 6, we get the following result.

THEOREM 4. *The schedule produced by algorithm schedule is completed strictly within* $(\frac{5}{4} + 2^{-k+1})OPT_{MM}(I)$.

Thus we have shown that the algorithm *schedule*, given enough iterations, produces a schedule that exceeds the optimal schedule by at most $\frac{5}{4}OPT_{MM}(I)$. Furthermore, if the jobs are sorted by their processing times, each iteration of binary search can be completed in linear time, thus resulting in $O(kn + n \log n)$ running time for **procedure** *schedule.*

In spite of the close relationship between the bin-packing and minimum makespan problems, the state of knowledge for these two problems is very different. For the bin-packing problem there exists a polynomial-time algorithm to find a feasible packing that uses at most $(1+\varepsilon)OPT_{BP}(I) + O(\varepsilon^{-2})$ machines, for every fixed $\varepsilon > 0$, where the running time is polynomial in $(1/\varepsilon)$ [KK]. For $P\|C_{\max}$, no polynomial-time approximation algorithm is known that achieves a better guarantee than $\frac{6}{5}OPT_{MM}(I)$. As a result, we believe that it is important to continue studying the relationship between these two problems, in the hope of achieving better guarantees for the minimum makespan problem.

6. Some computational experience. The algorithms presented in this paper are extremely easy to implement. Approximation algorithms usually perform significantly better than the performance guarantees indicate, and these algorithms are no exceptions to the rule. In the tables below we compare the performance of our algorithm for $P\|C_{\max}$ to the LPT and MULTIFIT algorithms.

For the first experiment, we considered instances where the processing times of the jobs were uniformly distributed in the interval $[0, 1]$. We ran 50 trials for each problem size. Since the optimal value of each instance was practically unobtainable, we measured the performance of each algorithm by computing the ratio

$$\frac{\text{cost of heuristic solution}}{\text{lower bound for optimal solution}}$$

where the lower bound was obtained by computing

$$\max\left\{\frac{1}{m}\left(\sum_{j=1}^{n} p_j\right) \max_j p_j\right\}.$$

This ratio is, of course, an upper bound on the performance of the algorithm. Table 1 demonstrates the dependence of the actual performance on the ratio n/m.

In Table 2 we see that there is little dependence on n if n/m remains constant.

It is also interesting to consider other distributions of processing times. For example, we approximated a normal distribution by considering processing times that were the average of ten uniformly distributed values; see Table 3.

Table 1

10-machine problems with uniform [0, 1] processing times

n	LPT	MULTIFIT	$\frac{3}{2}\varepsilon$
20	1.083	1.079	1.081
30	1.061	1.017	1.051
40	1.030	1.009	1.027
50	1.022	1.005	1.020
100	1.005	1.001	1.005
200	1.001	1.000	1.001

Table 2

Uniform processing times with $n/m = 3$

n	m	LTP	MULTIFIT	$\frac{3}{2}\varepsilon$
30	10	1.061	1.017	1.051
60	20	1.069	1.010	1.049
120	40	1.078	1.006	1.059

Table 3

10-machine problems with quasi-normal [0, 1] processing times

n	LTP	MULTIFIT	$\frac{3}{2}\varepsilon$
20	1.041	1.041	1.041
30	1.021	1.059	1.072
40	1.016	1.049	1.045
50	1.013	1.035	1.030
100	1.006	1.014	1.008
200	1.003	1.006	1.003

Finally (see Table 4) we consider instances where the lower bound used in computing the approximate performance of the algorithms is the optimal value as well. We do this by choosing a random schedule such that each machine is scheduled for exactly one unit of time. This is done by first randomly selecting the number of jobs n_j to be scheduled for machine j, between 2 and *range* + 1, where *range* is an input to

Table 4

10-machine problems with known optimal schedule

range	LTP	MULTIFIT	$\frac{3}{2}\varepsilon$
2	1.042	1.012	1.034
3	1.028	1.008	1.024
4	1.023	1.006	1.022
5	1.014	1.004	1.012
10	1.007	1.001	1.006
20	1.002	1.000	1.002

the procedure. The processing times are then generated by randomly dividing the unit of time for machine j into n_j parts. Note that the number of jobs for an instance is not fixed, but the expected number is $((range+3)/2)m$.

Acknowledgments. We are grateful to Richard Karp, who made several useful suggestions about a preliminary version of these results, and to the anonymous referee, whose comments helped to tighten the bound for this algorithm.

REFERENCES

[CGJ] E. G. Coffman, Jr., M. R. Garey and D. S. Johnson (1978), *An application of bin-packing to multiprocessor scheduling*, SIAM J. Comput., 7, pp. 1-17.

[Fs] D. K. Friesen (1978), *Sensitivity analysis for heuristic algorithms*, Technical Report UIUCDCS-R-78-939, Dept. Computer Science, Univ. Illinois, Champaign-Urbana.

[Ga] M. Gardner (1979), *Mathematical games: some packing problems that cannot be solved by sitting on the suitcase*, Scientific American 241, October 1979.

[GJ] M. R. Garey and D. S. Johnson (1979), *Computers and Intractability: A Guide to the Theory of* NP-*completeness*, Freeman, San Francisco.

[GLLR] R. L. Graham, E. L. Lawler, J. K. Lenstra and A. H. G. Rinnooy Kan (1979), *Optimization and approximation in deterministic sequencing and scheduling: a survey*, Ann. Disc. Math., 5, pp. 287-326.

[Gr] R. L. Graham (1966), *Bounds for certain multiprocessing anomalies*, Bell System Tech. J., 45, pp. 1563-1581.

[Gr1] ———, (1969), *Bounds on multiprocessing timing anomalies*, SIAM J. Appl. Math., 17, pp. 263-269.

[KK] N. Karmarkar and R. M. Karp (1982), *An efficient approximation scheme for the one-dimensional bin-packing problem*, 23rd Symposium on Foundations of Computer Science, pp. 312-320.

[PS] C. H. Papadimitriou and K. Steiglitz (1982), *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, NJ.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

010

SIMPLIFIED RELIABILITIES FOR CONSECUTIVE-k-OUT-OF-n SYSTEMS*

F. K. HWANG†

Abstract. Reliabilities for consecutive-k-out-of-n systems are typically given in the form of recursive equations. Some attempts have been made to use combinatorics to obtain closed-form solutions, but the solutions contain $k-1$ summations. In this paper we obtain closed form solutions with one summation over n/k terms. For $k=2$ we are able to eliminate all summations. We apply our result to compute the reliability of a k-loop computer network.

Key words. consecutive-k-out-of-n, reliability

AMS(MOS) subject classifications. 60C05, 90B25, 62N05

1. Introduction. A consecutive-k-out-of-n F system is usually defined as a system of n components in sequence where the system fails if and only if k consecutive components fail. The definition has been extended to cover a system of n components in cycle. We will refer to one as a *consecutive-k line* and the other as a *consecutive-k cycle.*

We consider the case that component i has probability p_i of working and probability $q_i=1-p_i$ of failing and the states of the components are independent. The special case $p_i=p$ is of practical importance and has received the most attention. We will refer to the $p_i=p$ case as the *Bernoulli* model.

For given $p_1, \cdots, p_n$, the *reliability* of a k-consecutive line or cycle is the probability that the system works, i.e., that there do not exist k consecutive failed components. Computing reliabilities for such systems and setting bounds for them have recently aroused a great deal of interest [1], [2], [3], [5], [8], [9], [12]. Most of the computing algorithms proposed are recursive in n. The fastest algorithm for the consecutive-k line requires $O(n)$ time [8], [12], and the fastest algorithm for the consecutive-k cycle requires $O(nk^2)$ time ($O(nk)$ time for the Bernoulli model) [8]. Although those recursive algorithms are computationally efficient, they have the usual disadvantage associated with a recursive algorithm of being a black box grinding out only numbers. The dependence of the reliability on the system parameters is hidden in the equations.

For the Bernoulli model reliabilities can be computed by using a combinatorial approach which is more explicit in nature. Let $N_L(j, n, k)(N_C(j, n, k))$ denote the number of ways of having working consecutive-k lines (cycles) conditional on j failed components. Since each such line (cycle) is associated with the probability q^jp^{n-j}, we have the reliability (x is either L or C)

$$R_x(p, k, n)=\sum_j q^jp^{n-j}N_x(j, n, k).$$

(Since we are dealing with a physical problem here, it is natural to define the binomial coefficient $\binom{y}{z}=0$ for negative y. With this definition sometimes we can avoid writing a messy expression of the upper range of a sum.) Several combinatorial arguments have been given to compute $N_L(j, n, k)$ without much success. In general $N_L(j, n, k)$ has been given in a closed form involving $(k-2)$ summations. In this paper we show that $N_L(j, n, k)$ can be given in a closed form involving only one summation over n/k terms. We then use this result to give $R_L(p, k, n)$ and $R_C(p, k, n)$ in closed forms also with one summation over n/k terms. For $k=2$ we are able to reduce $R_L(p, 2, n)$ and

* Received by the editors November 8, 1984, and in revised form May 21, 1985.

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

$R_C(p, 2, n)$ to closed forms without any summations. Finally we apply our results to compute the reliability of a k-loop network.

2. Some combinatorial results. Chiang and Niu [3] showed that

$$N_L(j, n, 2) = \binom{n-j+1}{j}$$

by considering the number of ways of placing at least one working component between every two failed components. Derman, Lieberman and Ross [5] pointed out that this approach does not seem to generalize beyond $k=2$. Instead, they considered the number of ways of placing at most $k-1$ failed components between every two working components and obtained the recursive equation

$$N_L(j, n, k) = \sum_{i \geqq 0} \binom{n-j+1}{i} N_L(j-(k-1)i, n-ki, k-1) \quad \text{for } k \geqq 3.$$

(Our N function is different from their N function which has the parameter $r = n-j+1$ instead of n.) By substituting the solution of $N(j, n, 2)$ into the recursive equation for $N(j, n, 3)$, we have

$$N_L(j, n, 3) = \sum_{i \geqq 0} \binom{n-j+1}{i} \binom{n-j-i+1}{j-2i}.$$

One can keep on substituting to obtain a closed-form solution of $N(j, n, k)$ involving $k-2$ summations. Bolinger [1] gave the recursive equations

$$N_L(j, n, k) = \binom{n}{j} \quad \text{for } j < k,$$

$$N_L(j, j, k) = 1 \quad \text{for } j \geqq k,$$

$$N_L(j, n, k) = \sum_{i=0}^{k-1} N_L(j-i, n-1-i, k) \quad \text{for } n > j \geqq k$$

by considering the position of the first working component. He tabulated numerical values for $N_L(j, n, 3)$ for $j \leqq 9$ and $n \leqq 10$.

We now show that the recursive equations of Bolinger can be solved in closed form with one summation over n/k terms.

THEOREM 1.

$$N_L(j, n, k) = \sum_{i \geqq 0} (-1)^i \binom{n-j+1}{i} \binom{n-ki}{n-j} \quad \textit{for } n > j \geqq k \geqq 1.$$

Proof. We prove this by substituting the solution into Bolinger's recursive equations.

$$\begin{aligned}
&\sum_{l \geqq 0} N_L(j-l, n-1-l, k) \\
&= \sum_{l=0}^{k-1} \sum_{i \geqq 0} (-1)^i \binom{n-j}{i} \binom{n-1-l-ki}{n-1-j} \\
&= \sum_{i \geqq 0} (-1)^i \binom{n-j}{i} \sum_{l=0}^{k-1} \binom{n-1-l-ki}{n-1-j} \\
&= \sum_{i \geqq 0} (-1)^i \binom{n-j}{i} \sum_{x=n-k-ki}^{n-1-ki} \binom{x}{n-1-j}
\end{aligned}$$

$$= \sum_{i\geqq 0} (-1)^i \binom{n-j}{i}\left[\binom{n-ki}{n-j} - \binom{n-k-ki}{n-j}\right]$$

(by using combinatorial identity (1.51) in Gould [6])

$$= \sum_{i\geqq 0} (-1)^i \left[\binom{n-j}{i} + \binom{n-j}{i-1}\right]\binom{n-ki}{n-j}$$

$$= \sum_{i\geqq 0} (-1)^i \binom{n-j+1}{i}\binom{n-ki}{n-j} = N_L(j, n, k).$$

COROLLARY.

$$N_L(j, n, 1) = 0 \quad \textit{except } N_L(0, n, 1) = 1.$$

Proof.

$$\sum_{i\geqq 0} (-1)^i \binom{n-j+1}{i}\binom{n-i}{n-j} = \begin{cases} 1 & \text{for } j = 0, \\ \binom{j-1}{-1} = 0 & \text{for } j > 0, \end{cases}$$

by using the combinatorial identity (3.49) in Gould [6].

Theorem 1 can be directly argued. Interpret $n-j+1$ as the number of spaces between the $n-j$ working components (including the two ends) into which the failed components are to be placed. Suppose that i of these spaces contain k or more failed components. We can count the number of such combinations by eliminating ik failed components and counting the number of ways of selecting $n-j$ working components from a total of now $n-ki$ components. An inclusion-exclusion argument then yields Theorem 1.

Since $N_L(j, n, k)$ as given in Theorem 1 must also solve the recursive equations of Derman, Lieberman and Ross, we have the following result.

THEOREM 2.

$$\sum_{i\geqq 0} (-1)^i \binom{n-j+1}{i}\binom{n-ki}{n-j} = \sum_{l>0} \binom{n-j+1}{l} \sum_{i\geqq 0} (-1)^i \binom{n-j-l+1}{i}\binom{n-kl-ki}{n-j-l}$$

$$= \sum_{i>0} (-1)^i \sum_{l\geqq 0} \binom{n-j+1}{l}\binom{n-j-l+1}{i}\binom{n-kl-ki}{n-j-l}.$$

In particular, for $k=3$ we have

COROLLARY.

$$\sum_{i\geqq 0} (-1)^i \binom{n-j+1}{i}\binom{n-3i}{n-j} = \sum_{i\geqq 0} \binom{n-j+1}{i}\binom{n-j-i+1}{j-2i}.$$

These combinatorial identities seem to be new as they are not listed in Gould.

3. The reliability.

THEOREM 3.

$$R_L(p, k, n) = \sum_{i\geqq 0} (-1)^i p^{i-1} q^{ki}\left[\binom{n-ki+1}{i} - q\binom{n-ki}{i}\right] \quad \textit{for } k \geqq 1.$$

Proof. From Theorem 1 we have

$$R_L(p,k,n)=\sum_{j\geqq 0} q^j p^{n-j}\sum_{i\geqq 0}(-1)^i\binom{n-j+1}{i}\binom{n-ki}{n-j}$$

$$=\sum_{i\geqq 0}(-1)^i\frac{(n-ki)!}{i![n-(k+1)i+1]!}\sum_{j\geqq 0}q^jp^{n-j}(n-j+1)\binom{n-(k+1)i+1}{j-ki}$$

$$=\sum_{i\geqq 0}(-1)^i\frac{(n-ki)!}{i![n-(k+1)i+1]!}q^{ki}p^{i-1}$$

$$\cdot\sum_{l\geqq 0}q^lp^{n-(k+1)i+1-l}(n-ki+1-l)\binom{n-(k+1)i+1}{l}$$

$$=\sum_{i\geqq 0}(-1)^i\frac{(n-ki)!}{i![n-(k+1)i+1]!}q^{ki}p^{i-1}$$

$$\cdot\sum_{l=0}^{n-(k+1)i+1}q^lp^{n-(k+1)i+1-l}(n-ki+1-l)\binom{n-(k+1)i+1}{l}$$

$$=\sum_{i\geqq 0}(-1)^i\frac{(n-ki)!}{i![n-(k+1)i+1]!}q^{ki}p^{i-1}[n-ki+1-\{n-(k+1)i+1\}q]$$

$$=\sum_{i\geqq 0}(-1)^ip^{i-1}q^{ki}\left[\binom{n-ki+1}{i}-q\binom{n-ki}{i}\right].$$

Derman, Lieberman and Ross proved that

$$R_C(p,k,n)=p^2\sum_{l=0}^{k-1}(l+1)q^lR_L(p,k,n-l-2).$$

So we can also obtain $R_C(p,k,n)$ from Theorem 3. We now show that $R_C(p,k,n)$ can also be reduced to a sum of n/k terms. First a combinatorial lemma.

LEMMA 1.

$$f(n)\equiv\sum_{i\geqq 0}(-pq^k)^i\left[\binom{n-ki}{i}-p\sum_{l=0}^{k-1}\binom{n-1-ki-l}{i}q^l\right]=q^n \quad \textit{for } 0\leqq k\leqq n.$$

Proof. Lemma 1 is trivially true for $n=0$. We prove the general case by induction on n. For $k=0$

$$f(n)=\sum_{i\geqq 0}(-p)^i\binom{n}{i}=(1-p)^n=q^n.$$

For $k=n$

$$f(n)=\sum_{i\geqq 0}(-pq^n)^i\left[\binom{n-ni}{i}-p\sum_{l=0}^{n-1}\binom{n-1-ni-l}{i}q^l\right]$$

$$=1-p\sum_{l=0}^{n-1}q^l=q^n.$$

For $1\leqq k\leqq n-1$

$$f(n)-f(n-1)=\sum_{i\geqq 1}(-pq^k)^i\left[\binom{n-1-ki}{i-1}-p\sum_{l=0}^{k-1}\binom{n-2-ki-l}{i-1}q^l\right]$$

$$=-pq^k\sum_{i\geqq 0}(-pq^k)^i\left[\binom{n-1-k-ki}{i}-p\sum_{l=0}^{k-1}\binom{n-2-k-ki-l}{i}q^l\right]$$

$$=-pq^kq^{n-1-k}=-pq^{n-1}\quad\text{by induction.}$$

Hence

$$f(n)=f(n-1)-pq^{n-1}=q^{n-1}-pq^{n-1}=q^n \quad \text{by induction.}$$

The fact that $f(n)$ is independent of k is rather surprising. Lemma 1 is also a very rich combinatorial identity so that some of its special cases seem to be new. For example, with $k=1$ we have

COROLLARY.

$$\sum_{i\geqq 0}(-pq)^i\left[\binom{n-1}{i}-p\binom{n-1-i}{i}\right]=q^n.$$

THEOREM 4.

$$R_C(p,k,n)=\sum_{i\geqq 0}(-pq^k)^i\binom{n-ki}{i}-q^n+k\sum_{i\geqq 0}(-pq^k)^{i+1}\binom{n-k(i+1)-1}{i}.$$

Proof.

$$\begin{aligned}R_C(p,k,n)&=p^2\sum_{l=0}^{k-1}(l+1)q^lR_L(p,k,n-l-2)\\&=p^2\sum_{l=i}^{k-1}(l+1)q^l\sum_{i\geqq 0}(-1)^ip^{i-1}q^{ki}\left[\binom{n-l-1-ki}{i}-q\binom{n-l-2-ki}{i}\right.\\&=p\sum_{i\geqq 0}(-pq^k)^i\sum_{l=0}^{k-1}(l+1)q^l\left[\binom{n-1-ki-l}{i}-q\binom{n-2-ki-l}{i}\right]\\&=p\sum_{i\geqq 0}(-pq^k)^i\left[\sum_{l=0}^{k-1}\binom{n-1-ki-l}{i}q^l-k\binom{n-k-1-ki}{i}q^k\right]\\&=\sum_{i\geqq 0}(-pq^k)^i\binom{n-ki}{i}-q^n+k\sum_{i\geqq 0}(-pqk)^{i+1}\binom{n-k(i+1)-1}{i}\end{aligned}$$

by Lemma 1.

We now show that $R_L(p,2,n)$ and $R_C(p,2,n)$ can be given without any summation.

THEOREM 5.

$$R_L(p,2,n)=p^n\frac{(2q/p+1+\sqrt{1+4q/p})^{n+2}+(-1)^{n+1}(2q/p)^{n+2}}{2^{(n+1)/2}(2q/p+1+\sqrt{1+4q/p})^{(n+1)/2}(4q/p+1+\sqrt{1+4q/p})},$$

$$R_C(p,2,n)=np^n\frac{(1+\sqrt{1+4q/p})^n+(1-\sqrt{1+4q/p})^n}{2^n}.$$

Proof.

$$\begin{aligned}R_L(p,2,n)&=\sum_{j=0}^{\lceil n/2\rceil}q^jp^{n-j}N(j,n,2)\\&=\sum_{j=0}^{\lceil n/2\rceil}q^jp^{n-j}\binom{n+1-j}{j}\\&=p^n\sum_{j=0}^{\lceil n/2\rceil}\binom{n+1-j}{j}\left(\frac{q}{p}\right)^j\\&=p^n\frac{(2q/p+1+\sqrt{1+4q/p})^{n+2}+(-1)^{n+1}(2q/p)^{n+2}}{2^{(n+1)/2}(2q/p+1+\sqrt{1+4q/p})^{(n+1)/2}(4q/p+1+\sqrt{1+4q/p})}\end{aligned}$$

by using combinatorial identity (1.70) in Gould.

The number of ways of selecting j objects, $j \geqq 1$, from n objects arranged in a cycle without two consecutive objects both being selected was given by David and Barton [4] to be 1 for $n=1$ and to be

$$\frac{n}{n-j}\binom{n-j}{j} \quad \text{for } n \geqq 2.$$

Therefore for $n \geqq 2$

$$\begin{aligned} R_C(p,2,n) &= \sum_{j=0}^{\lfloor n/2 \rfloor} q^j p^{n-j} \binom{n-j}{j} \frac{n}{n-j} \\ &= np^n \sum_{i=0}^{\lfloor n/2 \rfloor} \binom{n-j}{j} \frac{1}{n-j} \left(\frac{q}{p}\right)^j \\ &= \frac{np^n (1+\sqrt{1+4q/p})^n + (1-\sqrt{1+4q/p})^n}{2^n} \end{aligned}$$

by using the combinatorial identity (1.64) in Gould.

4. An application to computer networks. Consider n stations denoted by the residues modulo n. A k-loop network for n stations consists of k loops with links $i \to i+s_j$, $i=0, 1, \cdots, n-1$, $j=1, \cdots, k$. Clearly if the network is connected, we can always assume $s_1 = 1$. k-loop networks have been widely studied [7], [10], [11], [13] as topologies for computer networks.

Assume that each station can fail independently with probability q but links always work. A common measure for reliability is to call the system failed if there exist two working stations A and B such that every path from A to B must go through a station which has failed. We now use the result of § 3 to compute the reliability of a k-loop network with $s_i = i$.

LEMMA 2. *The network fails if and only if it contains at least two working stations and k consecutive failing stations.*

Proof. By our definition of network failure, clearly a network can fail only if at least two stations are working. In the rest of the proof we assume the existence of two working stations.

Suppose that there exist k consecutive failed stations $i, i+1, \cdots, i+k-1$. Let j be the first working station in the sequence $i-1, i-2, \cdots$. Let $k \neq j$ be another working station. Note that j and k must exist by our assumption of the existence of two working stations. Since every path of j must go through one of the k stations, $j+1, j+2, \cdots, j+k-1$, all failing, j cannot reach k.

Next suppose that there do not exist k consecutive failed stations. Then a working station i can always reach the first working station in the sequence $i+1, i+2, \cdots, i+k-1$ and by our assumption, at least one of them is working. But the fact that every working station can reach its next working station implies that all working stations can reach each other. The lemma is proved.

Therefore the reliability of the network is simply the reliability of a consecutive-k cycle plus the probability that at most one station works. We have

THEOREM 6. *The reliability of a k-loop network with $s_i = i$ and n stations is*

$$q^{n-1} + \sum_{i \geqq 0} (-pq^k)^i \binom{n-ki}{i} - q^n + k \sum_{i \geqq 0} (-pq^k)^{i+1} \binom{n-k(i+1)-1}{i} \quad \textit{for } k \geqq 1.$$

REFERENCES

[1] R. C. Bolinger, *Direct computation for consecutive-k-out-of-n: F systems*, IEEE Trans. Reliability, R-31 (1982), pp. 444-446.

[2] R. C. Bolinger and A. A. Salvia, *Consecutive-k-out-of-n: F networks*, IEEE Trans. Reliability, R-31 (1982), pp. 53-56.

[3] D. C. Chiang and S. C. Niu, *Reliability of consecutive-k-out-of-n: F systems*, IEEE Trans. Reliability, R-30 (1981), pp. 87-89.

[4] F. N. David and D. E. Barton, *Combinatorial Chances*, Hafner, New York, 1962.

[5] C. Derman, G. J. Lieberman and S. M. Ross, *On the consecutive-k-out-of-n system*, IEEE Trans. Reliability, R-31 (1982), pp. 57-63.

[6] H. W. Gould, *Combinatorial Identities*, revised edition, Morgantown, WV, 1972.

[7] A. Grnarov, L. Kleinrock and M. Gerla, *A highly reliable distributed loop network architecture*, Proc. 1980 International Symposium on Fault-Tolerant Computing, Kyoto, October, 1980, pp. 319-324.

[8] F. K. Hwang, *Fast solutions for consecutive-k-out-of-n system*, IEEE Trans. Reliability, R-31 (1982), pp. 447-448.

[9] J. M. Kontoleon, *Reliability determination of a r-successively-out-of-n: F system*, IEEE Trans. Reliability, R-29 (1980), p. 435.

[10] M. T. Liu, *Distributed loop computer network*, in Advances in Computers, Vol. 17, Academic Press, New York, 1978, pp. 163-221.

[11] C. S. Raghavendra and M. Gerla, *Optimal loop topologies for distributed systems*, Proc. 7th Data Communication Symposium, Mexico City, October, 1981, pp. 218-223.

[12] F. G. Shanthikumar, *Recursive algorithm to evaluate the reliability of a consecutive-k-out-of-n: F system*, IEEE Trans. Reliability, R-32 (1982), pp. 442-443.

[13] C. K. Wang and D. Coppersmith, *A combinatorial problem related to multimodule memory organizations*, J. Assoc. Comput. Mach., 21 (1974), pp. 392-402.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

011

ON THE SPECTRAL RADIUS OF COMPLEMENTARY ACYCLIC MATRICES OF ZEROS AND ONES*

RICHARD A. BRUALDI† AND ERNIE S. SOLHEID†

Abstract. For an $n \times n$ complementary acyclic matrix A of 0's and 1's we show that the spectral radius $\rho(A)$ of A satisfies $\rho(A) \geqq n-2$ and determine those matrices A for which equality holds. When A is an $n \times n$ irreducible, complementary tree matrix, we also obtain that $\rho(A) \leqq \rho_n$, where ρ_n is the largest root of the polynomial $\lambda^3-(n-2)\lambda^2-(n-3)\lambda-1$.

Key words. spectral radius, complementary acyclic matrix, tree

1. Introduction. Let $A=[a_{ij}]$ be an $m \times n$ (0, 1)-matrix, that is, a matrix all of whose entries are 0's and 1's. We associate with A a *bipartite graph* $G_0(A)$ whose edges correspond to the 0's of A. The graph $G_0(A)$ has $m+n$ vertices, $x_1, \cdots, x_m$ (*row vertices*) and $y_1, \cdots, y_n$ (*column vertices*), where there is an edge between x_i and y_j if and only if $a_{ij}=0$ $(1 \leqq i \leqq m, 1 \leqq j \leqq n)$. The matrix A is called *complementary acyclic* [3] if the graph $G_0(A)$ has no cycles. When $G_0(A)$ is connected, that is, $G_0(A)$ is a tree, we say that A is a *complementary tree matrix*. Thus the complementary acyclic matrix A has at most $m+n-1$ 0's with equality if and only if it is a complementary tree matrix.

Now let $B=[b_{ij}]$ be an $n \times n$ nonnegative matrix. Below we summarize those parts of the Perron-Frobenius theory [1], [6] that we require. The *spectral radius* $\rho(B)$ of B is the maximum absolute value of an eigenvalue of B. The matrix B is called *reducible* if there exists a permutation matrix P such that

$$P^tAP=\begin{bmatrix} A_1 & 0 \\ A_{21} & A_2 \end{bmatrix}$$

where A_1 and A_2 are square, nonvacuous matrices; B is *irreducible* when it is not reducible. If $C=[c_{ij}]$ is another $n \times n$ nonnegative matrix, we write $B \leqq C$ when $b_{ij} \leqq c_{ij}$ for all i and j. Then the following hold:

(1.1) $\rho(B)$ is an eigenvalue of B and has an associated nonnegative eigenvector u (when B is irreducible, u is positive).

(1.2) Let the row sums of B be $r_1, \cdots, r_n$. Then

$$\min\{r_1, \cdots, r_n\} \leqq \rho(B) \leqq \max\{r_1, \cdots, r_n\}.$$

When B is irreducible and not all row sums are equal, both of the inequalities are strict.

(1.3) Let z be a positive vector. If $Bz \geqq rz$ (respectively, $Bz \leqq rz$), then $\rho(B) \geqq r$ (respectively, $\rho(B) \leqq r$) with equality for irreducible B if and only if $Bz=rz$.

(1.4) Let $B \leqq C$. Then $\rho(B) \leqq \rho(C)$ with strict inequality when C is irreducible and $B \neq C$.

(1.5) If B' is a proper principal submatrix of the irreducible matrix B, then $\rho(B') < \rho(B)$.

* Received by the editors January 2, 1985, and in revised form, June 17, 1985. This research was partially supported by the National Science Foundation under grant DMS-8320189.

† Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706.

Let $\mathfrak{U}_n$ be the set of all $n \times n$ $(0,1)$-matrices, and let $\mathcal{P} \subseteq \mathfrak{U}_n$. We can formulate the following general

Problem. Determine

$$\tilde{\rho} = \min \{\rho(A) \colon A \in \mathcal{P}\} \quad \text{and} \quad \bar{\rho} = \max \{\rho(A) \colon A \in \mathcal{P}\}.$$

It is of interest to find sets $\mathcal{P}$ for which $\tilde{\rho}$ or $\bar{\rho}$ can be determined. For if $A \in \mathcal{P}$, then $\tilde{\rho} \leqq \rho(A) \leqq \bar{\rho}$, and these bounds may improve on the bound in (1.2) or other known bounds for the spectral radius. In [2] and [5] the problem of determining $\bar{\rho}$ was considered when $\mathcal{P}$ is the set of $n \times n$ $(0,1)$-matrices with a specified number of 1's, and when $\mathcal{P}$ is the set of symmetric $n \times n$ $(0,1)$-matrices with zero trace and a specified number of 1's. In this note we solve the above problem when $\mathcal{P}$ is the set of $n \times n$ complementary acyclic $(0,1)$-matrices and when $\mathcal{P}$ is the set of (irreducible) $n \times n$ complementary tree matrices.

Our work can be viewed in two ways: (1) as a contribution to the spectral theory of graphs [4], which is often useful in proving theorems about graphs, and (2) as a contribution to the highly developed theory of nonnegative matrices and spectral radius [1], [6], which has been applied in many diverse fields such as economics, probability, and demography.

2. A lower bound. We begin with the following lemma which is a special case of a method of B. Schwarz [7].

LEMMA 2.1. *Let $A = [a_{ij}]$ be an $n \times n$ irreducible $(0,1)$-matrix and let $z = (z_1, \cdots, z_n)^t$ be a positive eigenvector corresponding to the spectral radius ρ of A. Suppose that for some i there exists j and k with $j < k$ such that $a_{ij} = 0$ and $a_{ik} = 1$. Let B be the matrix obtained from A by interchanging the entries a_{ij} and a_{ik}.*

(i) *If $z_1 \leqq \cdots \leqq z_n$, then $\rho(B) \leqq \rho(A)$.*

(ii) *If $z_1 \geqq \cdots \geqq z_n$, then $\rho(B) \geqq \rho(A)$.*

Proof. In case (i) we have $Bz \leqq Az = \rho z$ while in case (ii) we have $Bz \geqq Az = \rho z$. The result now follows from (1.3).

We remark that the conclusions of Lemma 2.1 hold when B is obtained from A by several interchanges of entries as long as in each interchange the 0 precedes the 1.

THEOREM 2.2. *Let $n \geqq 3$ and let A be an $n \times n$ complementary acyclic $(0,1)$-matrix. Then*

$$\rho(A) \geqq n - 2. \tag{2.1}$$

For A irreducible, equality holds if and only if there exists a permutation matrix P such that P^tAP has the form

$$\begin{bmatrix} A_1 & T \\ J & A_2 \end{bmatrix}, \tag{2.2}$$

where A_1 is a square nonvacuous matrix with exactly one 0 in each column, A_2 is a square nonvacuous matrix with exactly one 0 in each row, T is a complementary tree matrix, and J is a matrix of all 1's.

For A reducible, equality holds if and only if there is a permutation matrix P such that P^tAP has one of the forms

$$\left[\begin{array}{c|c} & 0 \quad \cdots \quad 0 \\ \cline{2-2} u & C \\ & \end{array}\right] \tag{2.3}$$

where C is an $(n-1) \times (n-1)$ matrix with exactly one 0 in each row and u has at most one 0, or the transpose of this form;

(2.4) *a complementary acyclic matrix of the form*

$$\left[\begin{array}{c|c|c} & \multicolumn{2}{c}{0 \quad \cdots \quad 0} \\ \cline{2-3} & \multicolumn{2}{c}{v} \\ \cline{2-3} u & 0 & \\ & \vdots & J \\ & 0 & \end{array}\right]$$

where J is an $(n-2)\times(n-2)$ matrix of 1*'s, or the transpose of this form.*

Proof. Let $A=[a_{ij}]$ be an $n\times n$ complementary acyclic matrix with minimum spectral radius $\rho=\rho(A)$. We first suppose that A is irreducible. Suppose A were not a complementary tree matrix. Then there exists a complementary tree matrix A' with $A'\leqq A$ and $A\neq A'$. By (1.4) $\rho(A')<\rho(A)$ which is a contradiction. Hence A is a complementary tree matrix. Let $z=(z_1,\cdots,z_n)^t$ be a positive eigenvector of A for ρ. After simultaneous row and column permutations of A, we may assume that $z_1\leqq\cdots\leqq z_n$. Since $G_0(A)$ is a tree, for each $i=1,\cdots,n$ there is a unique path γ_i from row vertex x_i to column vertex y_n. Let the first edge of the path γ_i be the edge $\{x_i, y_{j_i}\}$ joining x_i and y_{j_i} $(i=1,\cdots,n)$. Let B be the matrix obtained from A by interchanging for each $i=1,\cdots,n$ the entries a_{ij_i} and a_{in}. It follows from Lemma 2.1 that $\rho(B)\leqq\rho(A)$ and hence $\rho(B)=\rho(A)$. The graph $G_0(B)$ is obtained from the tree $G_0(A)$ by deleting for each $i=1,\cdots,n$ the edge $\{x_i, y_{j_i}\}$ and inserting the edge $e_i=\{x_i, y_n\}$. For each $k=1,\cdots,n-1$ let the first edge on the path in $G_0(A)$ from y_k to y_n be the edge $f_k=\{y_k, x_{i_k}\}$. Then f_k is not the first edge of any of the paths $\gamma_1,\cdots,\gamma_n$ of $G_0(A)$, and hence for each $k=1,\cdots,n-1$ f_k is an edge of $G_0(B)$. Since $G_0(B)$ has exactly $2n-1$ edges, it now follows that $e_1,\cdots,e_n, f_1,\cdots,f_{n-1}$ are precisely the edges of $G_0(B)$ and hence $G_0(B)$ is a tree. Thus B is a complementary tree matrix of the form

$$B=\left[\begin{array}{c|c} & 0 \\ & 0 \\ B_1 & \vdots \\ & \vdots \\ & 0 \end{array}\right]=\left[\begin{array}{c|c} & 0 \\ B_2 & \vdots \\ & 0 \\ \hline & 0 \end{array}\right]$$

where B_1 has exactly one 0 in each column and B_2 is an $(n-1)\times(n-1)$ matrix. Then $\rho(B)=\rho(B_2)$. Since B_2 has at least $n-2$ 1's in each column, it follows from (1.2) that $\rho(B_2)\geqq n-2$. Hence $\rho(A)=\rho(B)=\rho(B_2)\geqq n-2$, and (2.1) holds for A irreducible.

Now suppose equality holds in (2.1). Then $\rho(A)=\rho(B_2)=n-2$. First assume B_2 is irreducible. By (1.2) B_2 has exactly one 0 in each column. Hence the last row of B_1 contains no 0's and the last row of A contains exactly one 0. Let z' be obtained from z by deleting its last coordinate. Since $Bz\leqq Az=(n-2)z$, it follows that

$$B_2z'\leqq(n-2)z' \tag{2.5}$$

with equality if and only if $z_{j_i}=z_n$ for $i=1,\cdots,n-1$. Since $\rho(B_2)=n-2$, (1.3) implies that we have equality in (2.5). Hence $z_{j_i}=z_n$ for $i=1,\cdots,n-1$. Let A' be the complementary tree matrix obtained from A by interchanging a_{nj_n} and a_{nn}. Then A' is irreducible and $A'z\leqq Az=(n-2)z$. Using (1.3), we conclude that $z_{j_n}=z_n$ also. Therefore $Bz=Az=(n-2)z$. Let $k+1$ be the minimum of $\{j_1,\cdots,j_n\}$. From the monotonicity of z we now conclude that $z_{k+1}=\cdots=z_n$. Let A be partitioned as

$$A=\left[\begin{array}{c|c} U & V \\ \hline W & Z \end{array}\right]=[Q_1|Q_2]$$

where U is $k\times k$ and Q_1 is $n\times k$. Suppose $\{x_i, y_p\}$ is an edge of one of the paths $\gamma_1, \cdots, \gamma_n$. Then it follows that $y_p = y_{j_i}$ and hence that $p = j_i \geqq k+1$. Therefore each of $\gamma_1, \cdots, \gamma_n$ is a path of $G_0(Q_2)$. Since $G_0(A)$ is a tree, each column of Q_2 has a 0, and it follows that $G_0(Q_2)$ is connected and hence a tree. Thus Q_2 has $2n-k-1$ 0's and since A has $2n-1$ 0's, Q_1 has k 0's. Since each column of Q_1 must also have a 0, each column of Q_1 has exactly one 0. Since the last row of A contains only one 0, in the $j_n \geqq k+1$ column, the last row of W contains no 0's. Since $z_{k+1} = \cdots = z_n$, we may repeat the preceding argument on the matrix obtained from A by permuting row n and any of rows $k+1, \cdots, n-1$. The matrix corresponding to B_2 remains irreducible, and we conclude that W has no 0's and each of the rows of Z contains exactly one 0. Since each column of Q_1 has exactly one 0, we now conclude that each column of U contains exactly one 0. It now follows that V has exactly $n-1$ 0's. Since $G_0(V)$ is an acyclic graph with n vertices and $n-1$ edges, $G_0(V)$ is a tree and hence V is a complementary tree matrix. Thus A has the form (2.2).

Now assume B_2 is reducible. Since B_2 is complementary acyclic, B_2 has a row or column all of whose off-diagonal entries are 0. Since B is complementary acyclic and has only 0's in its last column, we conclude *that there is an integer* p with $0 \leqq p \leqq n-2$ such that

$$B=\left[\begin{array}{c|c|c|c} \overbrace{\qquad J \qquad}^{p} & & J & \begin{matrix}0\\ \vdots\\ 0\end{matrix} \\ \hline 0\ \cdots\ 0 & 1 & 0\ \cdots\ 0 & 0 \\ \hline J & & J & \begin{matrix}0\\ \vdots\\ 0\end{matrix} \\ \hline 1\ \cdots\ 1 & & 1\ \cdots\ 1 & 0 \end{array}\right] \begin{array}{l} \left.\vphantom{\begin{matrix}0\\ \vdots\\ 0\end{matrix}}\right\} p \\ \\ \\ \\ \\ \end{array}.$$

Since A is irreducible, row $p+1$ of B must have been obtained from row $p+1$ of A by interchanging the entries in columns $p+1$ and n (i.e. $j_{p+1} = p+1$). Let $\bar{B}$ be the matrix obtained from A by interchanging the entries a_{ij_i} and a_{in} for each $i = 1, \cdots, n$ with $i \neq p+1$:

$$\bar{B}=\left[\begin{array}{c|c|c|c} \overbrace{\qquad J \qquad}^{p} & & J & \begin{matrix}0\\ \vdots\\ 0\end{matrix} \\ \hline 0\ \cdots\ 0 & 0 & 0\ \cdots\ 0 & 1 \\ \hline J & & J & \begin{matrix}0\\ \vdots\\ 0\end{matrix} \\ \hline 1\ \cdots\ 1 & & 1\ \cdots\ 1 & 0 \end{array}\right] \begin{array}{l} \left.\vphantom{\begin{matrix}0\\ \vdots\\ 0\end{matrix}}\right\} p \\ \\ \\ \\ \\ \end{array}.$$

By Lemma 2.1, $\rho(\bar{B})=n-2$. Let $n\geqq 4$. Since column $p+1$ of $\bar{B}$ must contain at least one 1, $\bar{B}$ is readily seen to be irreducible. Let C be the matrix obtained from $\bar{B}$ by *replacing* the 1 in the $(p+1, n)$ position with a 0. Then using (1.4) we obtain

$$n-2=\rho(C)<\rho(\bar{B})=n-2.$$

This contradiction shows that for $n\geqq 4$, B_2 cannot be reducible. The same conclusion is easily reached when $n=3$. Hence when A is irreducible and $\rho(A)=n-2$, (2.2) holds.

Now suppose that A is an $n\times n$ irreducible matrix satisfying (2.2) where A_1 is a $k\times k$ matrix with $1\leqq k\leqq n-1$. Then A is a complementary tree matrix and by (2.1), $\rho(A)\geqq n-2$. Let $z=(z_1,\cdots,z_n)^t$ be a positive eigenvector corresponding to $\rho=\rho(A)$. Partition z as (z', z'') where z' has length k. Then $Az=\rho z$ implies

$$Jz'+A_2z''=\rho z'',$$

so that

$$(\rho I_{n-k}-A_2)z''=\mathbf{c}$$

where $\mathbf{c}$ is a constant vector. By (1.5), $\rho(A_2)<\rho$ and hence $\rho I_{n-k}-A_2$ is an invertible matrix. Therefore

$$z''=(\rho I_{n-k}-A_2)^{-1}\mathbf{c}=\frac{1}{\rho}\left(I_{n-k}+\frac{1}{\rho}A_2+\frac{1}{\rho^2}A_2^2+\cdots\right)\mathbf{c}.$$

Since $\mathbf{c}$ is a constant vector and since A_2 has exactly one 0 in each row, we conclude that z'' is a constant vector. The form (2.2) implies that $G_0[A_2^T]$ is a tree. Hence for each $i=1,\cdots,n$, the unique path γ_i of $G_0(A)$ from x_i to y_n is a path of $G_0[A_2^T]$. As before let the first edge of the path γ_i be the edge $\{x_i, y_{j_i}\}$ $(i=1,\cdots,n)$. Then $k+1\leqq j_i\leqq n$ for $i=1,\cdots,n$. We again let B be the matrix obtained from A by interchanging for each $i=1,\cdots,n$ the entries a_{ij_i} and a_{in}. Then B is a complementary tree matrix. Moreover, since z'' is a constant vector, ρ is an eigenvalue of B and hence $\rho(B)\geqq\rho$. Since by (2.2) A has exactly one 0 in its last row,

$$B=\left[\begin{array}{ccc|c} & & & 0\\ & B_2 & & \vdots\\ & & & 0\\ \hline 1 & \cdots & 1 & 0\end{array}\right]$$

where $\rho(B)=\rho(B_2)$. Since B is a complementary tree matrix, B_2 has exactly one 0 in each column and hence $\rho(B_2)=n-2$. Thus

$$n-2=\rho(B_2)=\rho(B)\geqq\rho(A)\geqq n-2,$$

and we conclude that $\rho(A)=n-2$. This completes the proof of the theorem when A is irreducible.

We now give the proof when A is reducible. In this case there exists a permutation matrix P such that P^tAP or P^tA^tP has the form

$$\left[\begin{array}{c|ccc} & 0 & \cdots & 0\\ \cline{2-4} u & & & \\ & & C & \\ & & & \end{array}\right]. \tag{2.6}$$

Since A is complementary acyclic, C has at most one 0 in each row and hence its row sums are at least $n-2$. From (1.2) we obtain that $\rho(A)=\rho(C)\geqq n-2$, and (2.1) holds.

We now consider when equality holds in (2.1) for A reducible. First suppose that $\rho(A)=n-2$. Then there exists a permutation matrix P such that P^tAP or P^tA^tP has the form (2.6) where $\rho(C)=\rho(A)=n-2$. Since A is complementary acyclic, C has at most one 0 in each row. Assume C is irreducible. Then it follows from (1.2) that C has exactly one 0 in each row. Since we have now accounted for $2n-2$ 0's, u can have at most one 0. Hence (2.3) is satisfied. Now assume C is reducible. Then after simultaneous row and column permutations, (2.6) becomes

$$\left[\begin{array}{c|cc} & 0\ \cdots & 0 \\ \cline{2-3} & \multicolumn{2}{c}{v} \\ \cline{2-3} u & 0 & \multicolumn{1}{|c}{} \\ & \vdots & \multicolumn{1}{|c}{D} \\ & 0 & \multicolumn{1}{|c}{} \end{array}\right]. \tag{2.7}$$

Since A is complementary acyclic, D contains no 0's and (2.7) is of the form (2.4). Thus when $\rho(A)=n-2$, A satisfies (2.3) or (2.4) after simultaneous row and column permutations. The converse follows easily, and the proof of the theorem is now complete.

3. An upper bound. For $n\geqq 3$ there exists an $n\times n$ irreducible complementary acyclic (0, 1)-matrix satisfying (2.2). Hence the lower bound (2.1) in Theorem 2.2 cannot be improved under the additional assumption of irreducibility. This is in contrast to the situation encountered for the upper bound for the spectral radius of complementary tree matrices.

First note that for $n\times n$ complementary acyclic matrices the maximum spectral radius is clearly n.

THEOREM 3.1. *Let A be an $n\times n$ complementary tree matrix. Then $\rho(A)\leqq n-1$ with equality if and only if there exists a permutation matrix P such that P^tAP has the form*

$$\left[\begin{array}{c|ccc} 0 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & J & \\ 0 & & & \end{array}\right]. \tag{3.1}$$

Proof. Since A is a complementary tree matrix, A has at least one 0 in each row and hence by (1.2), $\rho(A)\leqq n-1$. Now suppose $\rho(A)=n-1$. If A were irreducible, then by (1.2) again, A has exactly one 0 in each row, and hence $n=1$ and (3.1) holds. Hence we may assume A is reducible. Since A is a complementary tree matrix, the rows and columns of A or A^t can be simultaneously permuted to give

$$\left[\begin{array}{c|c} 0 & \cdots\ 0 \\ \hline & A' \end{array}\right], \tag{3.2}$$

where A' is an $(n-1)\times(n-1)$ complementary acyclic matrix with $\rho(A')=\rho(A)=n-1$. Using (1.2) once more, we conclude that A' is a matrix of all 1's and hence (3.2) equals (3.1). Since the matrix (3.1) has spectral radius $(n-1)$, the theorem follows.

Let ρ_n be the maximum spectral radius of an $n\times n$ irreducible, complementary tree matrix. It follows from Theorem 3.1 that $\rho_n<n-1$ for $n\geqq 3$. We now determine an irreducible, complementary tree matrix whose spectral radius is ρ_n.

THEOREM 3.2. *For $n \geqq 3$, $\rho_n = \rho(A_n)$ where A_n is the $n \times n$ matrix*

$$\left[\begin{array}{ccccc|c}
0 & 0 & 1 & \cdots & 1 & 1\\
1 & 1 & 1 & \cdots & 1 & 0\\
\vdots & \vdots & \vdots & & \vdots & \vdots\\
1 & 1 & 1 & \cdots & 1 & 0\\ \hline
0 & 1 & 0 & \cdots & 0 & 0
\end{array}\right],$$

which has characteristic polynomial $\lambda^{n-3}(\lambda^3-(n-2)\lambda^2-(n-3)\lambda-1)$. Therefore ρ_n is the largest root of $\lambda^3-(n-2)\lambda^2-(n-3)\lambda-1$.

Proof. Let $A=[a_{ij}]$ be an irreducible, complementary tree matrix with $\rho(A)=\rho_n$. Let $z=(z_1,\cdots,z_n)^t$ be a positive vector such that $Az=\rho_n z$. After simultaneous row and column permutations, we may assume that $z_1 \geqq \cdots \geqq z_n$. Let $\{x_i, y_{j_i}\}$ be the first edge of the unique path γ_i from row vertex x_i to column vertex y_n in the tree $G_0(A)$, $i=1,\cdots,n$. Let B be obtained from A by interchanging the entries a_{ij_i} and a_{in} for each $i=1,\cdots,n$ except for $i=k$ which is determined as follows. If there is a t such that row t contains all 0's except for a 1 in the last column, then there is exactly one such row and we set k equal to t. Otherwise, since A is irreducible, there exists at least one t with $1 \leqq t \leqq n-1$ such that γ_t has length at least 2, and we choose k to be any such t. It follows as in the proof of Theorem 2.2 that B is a complementary tree matrix. Moreover the last column of B contains exactly one 1. Since a complementary acyclic matrix is reducible only when it has a row or column all of whose nondiagonal entries are 0, it follows from the choice of k that B is irreducible. Since $Bz \geqq Az = \rho_n z$, (1.3) implies that $\rho(B) \geqq \rho_n$ and hence $\rho(B)=\rho_n$. Replacing A in the above argument by B^t, we obtain an irreducible, complementary tree matrix C with $\rho(C)=\rho_n$ such that C has both a row and a column with exactly one 1. After simultaneous row and column permutations we may assume C has the form

$$\left[\begin{array}{c|c|c|c}
 & & & 1\\ \hline
J & & J & \begin{matrix}0\\ \vdots\\ 0\end{matrix}\\ \hline
0 \cdots 0 & 1 & 0 \cdots 0 & 0\\ \hline
J & & J & \begin{matrix}0\\ \vdots\\ 0\end{matrix}
\end{array}\right]
\begin{array}{l} \\ \\ \\ \\ p \qquad (p \neq q). \\ \\ \\ \\ \end{array}$$

In addition to the 0's displayed, q the matrix (3.4) has exactly two more 0's lying in row 1 and column q. We note that the case $p=1$ may occur, in which case $q=n$.

Up to simultaneous row and column permutations, (3.4) gives rise to 5 cases:

(i) $p=1, q=n$,

(ii) $p=2, q=1$,

(iii) $p=2, q=3$,

(iv) $p=n, q=1$,

(v) $p=n, q=2$.

Each of these cases has a number of subcases. We have verified that the maximum spectral radius occurs when $p=n$ and $q=2$ (case (v)), and that the matrix C has the form (3.3). We give an indication of our verification by comparing the spectral radius of (3.3) with that of the matrix

$$D=\left[\begin{array}{ccccc|c} 1 & 0 & 1 & \cdots & 1 & 1 \\ 0 & 1 & 1 & \cdots & 1 & 0 \\ 1 & 1 & 1 & \cdots & 1 & \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 1 & \cdots & 1 & 0 \\ \hline 1 & 0 & 0 & \cdots & 0 & 0 \end{array}\right]$$

which arises in case (iv) for $n \geqq 4$. The characteristic polynomials of C and D are, respectively, $\lambda^{n-3}c(\lambda)$ and $\lambda^{n-3}d(\lambda)$ where

$$c(\lambda)=\lambda^3-(n-2)\lambda^2-(n-3)\lambda-1$$

and

$$d(\lambda)=\lambda^3-(n-1)\lambda^2+(2n-5).$$

Let $f(\lambda)=c(\lambda)-d(\lambda)=\lambda^2-(n-3)\lambda-(2n-4)$. The polynomial $f(\lambda)$ has a root

$$\lambda_n=\frac{(n-3)+\sqrt{n^2+2n-7}}{2}$$

between $n-2$ and $n-1$, and it follows that $c(\lambda) \leqq d(\lambda)$ for $n-2 \leqq \lambda \leqq \lambda_n$. Also, $c(\lambda_n)=\sqrt{n^2+2n-7}-n>0$. Hence the spectral radii of C and D lie in the interval between $(n-2)$ and λ_n. It now follows that the spectral radius of C is greater than the spectral radius of D. The remainder of the verification is carried out using similar arguments.

REFERENCES

[1] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[2] R. A. Brualdi and A. J. Hoffman, *On the spectral radius of* (0, 1)*-matrices*, Linear Alg. Appl., to appear.

[3] R. A. Brualdi and E. S. Solheid, *Maximum determinants of complementary acyclic matrices of zeros and ones*, Discrete Math., to appear.

[4] M. Doob, D. Cvetcovic and H. Sachs, *Spectra of Graphs*, second ed., Academic Press, New York, 1982.

[5] S. Friedland, *The maximum eigenvalue of* 0–1 *matrices with prescribed number of* 1's, Linear Alg. Appl., to appear.

[6] F. R. Gantmacher, *The Theory of Matrices*, vol. 2, translated by K. A. Hirsch, Chelsea, New York, 1959.

[7] B. Schwarz, *Rearrangements of square matrices with non-negative elements*, Duke Math. J., 31 (1964), pp. 45–62.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

012

USING THE QR FACTORIZATION AND GROUP INVERSION TO COMPUTE, DIFFERENTIATE, AND ESTIMATE THE SENSITIVITY OF STATIONARY PROBABILITIES FOR MARKOV CHAINS*

GENE H. GOLUB† AND CARL D. MEYER, JR.‡

Abstract. For an n-state finite, homogeneous, ergodic Markov chain, with transition matrix $\mathbf{P}$ and stationary distribution $\boldsymbol{\pi}$ we assume that the entries of $\mathbf{P}$ are differentiable functions of a parameter t and we obtain an expression for $d\boldsymbol{\pi}/dt$. This expression is given in terms of the group inverse of $\mathbf{I}-\mathbf{P}$ and is used in a sensitivity analysis of $\boldsymbol{\pi}$. Finally, it is demonstrated how a $\mathbf{QR}$ factorization can be used to simultaneously compute the stationary distribution of an ergodic chain along with estimates which gauge the sensitivity of the stationary distribution to perturbations in the transition probabilities.

Key words. finite Markov chain, stationary distribution, group inversion, sensitivity analysis, $\mathbf{QR}$ factorization, condition number, derivatives of stationary probabilities

AMS(MOS) subject classifications. 60J10, 65F-15, 25, 35, 15A-09, 12, 51

1. Introduction. For an n-state finite, homogeneous, ergodic Markov chain with transition matrix $\mathbf{P}=[p_{ij}]$, the stationary distribution is the unique row vector $\boldsymbol{\pi}$ satisfying

$$\boldsymbol{\pi}\mathbf{P}=\boldsymbol{\pi}, \qquad \sum \pi_i = 1.$$

Letting $\mathbf{A}_{n\times n}$ and $\mathbf{e}_{n\times 1}$ denote the matrices $\mathbf{A}=\mathbf{I}-\mathbf{P}$ and $\mathbf{e}=[1,1,\cdots,1]^T$, the stationary distribution $\boldsymbol{\pi}$ can be characterized as the unique solution to the linear system of equations defined by

$$\boldsymbol{\pi}\mathbf{A}=\mathbf{0} \quad \text{and} \quad \boldsymbol{\pi}\mathbf{e}=1.$$

(See Kemeny and Snell [11] for an elementary exposition of finite ergodic chains.)

The theory of finite Markov chains has long been a fundamental tool in the analysis of social and biological phenomena. More recently the ideas embodied in Markov chain models along with the analysis of a stationary distribution have proven to be useful in applications which do not fall directly into the traditional Markov chain setting. Some of these applications include the analysis of queuing networks (Kaufman [7]), the analysis of compartmental ecological models (Funderlic and Mankin [5]), and least squares adjustment of geodedic networks (Brandt [1]). Recently, the behavior of the numerical solution of systems of nonlinear reaction-diffusion equations has been analyzed by making use of the stationary distribution of a finite Markov chain in conjunction with the concept of group matrix inversion (Galeone [5]).

An ergodic chain manifests itself in the transition matrix $\mathbf{P}$ which must be row stochastic and irreducible. Of central importance is the sensitivity of the stationary distribution $\boldsymbol{\pi}$ to perturbations in the transition probabilities in $\mathbf{P}$. The sensitivity of $\boldsymbol{\pi}$ is most easily gauged by considering the transition probabilities in $\mathbf{P}$ to be differentiable functions. One approach, adopted by Conlisk [3], Schweitzer [11], and Funderlic and Heath [4] is to examine partial derivatives $\partial\boldsymbol{\pi}/\partial p_{ij}$. Our strategy is to consider the

* Received by the editors February 13, 1985, and in revised form May 24, 1985.

† Computer Science Department, Stanford University, Stanford, California 94305. The work of this author was supported in part by the National Science Foundation under grant MCS-78-11985.

‡ Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205. The work of this author was supported in part by the National Science Foundation under grant MCS-82-19500.

transition probabilities $p_{ij}(t)$ as differentiable functions of a single parameter t and to study the stationary distribution $\boldsymbol{\pi}(t)$ as a function of t. We present a new and very simple formulation for the derivative, $d\boldsymbol{\pi}(t)/dt$, of the stationary distribution directly in terms of the derivatives $dp_{ij}(t)/dt$ and entries from $\boldsymbol{\pi}(t)$ and a matrix $\mathbf{A}^{\#}(t)$, called the group inverse of $\mathbf{A}(t)=\mathbf{I}-\mathbf{P}(t)$. After the derivative $d\boldsymbol{\pi}(t)/dt$ has been obtained, we demonstrate its applicability by using it to deduce the relative sensitivity of a discrete Markov chain. This is followed by a first order perturbation analysis. Finally, it is demonstrated how a **QR** factorization can be used to simultaneously compute $\boldsymbol{\pi}$ along with estimates which gauge the sensitivity of $\boldsymbol{\pi}$ to perturbations in $\mathbf{P}$.

2. Background material. In this paper, we take advantage of results which are phrased in terms of the group inverse $\mathbf{A}^{\#}$ of $\mathbf{A}=\mathbf{I}-\mathbf{P}$. Below is a short summary concerning the matrix $\mathbf{A}^{\#}$. Proofs and additional background material on $\mathbf{A}^{\#}$ may be found in Campbell and Meyer [2] and Meyer [9], [10].

Background material concerning $\mathbf{A}^{\#}$.

(2.1) Each finite Markov chain has the property that $\mathbf{A}=\mathbf{I}-\mathbf{P}$ belongs to some multiplicative matrix group. ($\mathbf{P}$ is the transition matrix.) Let $\mathcal{G}$ denote the maximal subgroup containing $\mathbf{A}$. The inverse of $\mathbf{A}$ with respect to $\mathcal{G}$ is denoted by $\mathbf{A}^{\#}$ and the identity element in $\mathcal{G}$ is denoted by $\mathbf{E}$.

(2.2) For all finite Markov chains, the limiting matrix is the difference of the two identities $\mathbf{I}$ and $\mathbf{E}$ in the sense that

$$\mathbf{P}^{\infty}=\lim_{k\to\infty}\frac{\mathbf{I}+\mathbf{P}+\mathbf{P}^2+\cdots+\mathbf{P}^{k-1}}{k}=\mathbf{I}-\mathbf{E}=\mathbf{I}-\mathbf{A}\mathbf{A}^{\#}.$$

Of course, if the chain has a limiting matrix in the strong sense, then

$$\mathbf{P}^{\infty}=\lim_{k\to\infty}\mathbf{P}^k=\mathbf{I}-\mathbf{E}.$$

(2.3) If the chain is ergodic (i.e., $\mathbf{P}$ is irreducible), then

$$\mathbf{P}^{\infty}=\mathbf{I}-\mathbf{E}=\mathbf{I}-\mathbf{A}\mathbf{A}^{\#}=\mathbf{e}\boldsymbol{\pi}$$

where $\mathbf{e}$ is a column of 1's.

(2.4) The group inverse $\mathbf{A}^{\#}$ of $\mathbf{A}$ can be characterized as the unique matrix satisfying the three equations $\mathbf{A}\mathbf{A}^{\#}\mathbf{A}=\mathbf{A}$, $\mathbf{A}^{\#}\mathbf{A}\mathbf{A}^{\#}=\mathbf{A}^{\#}$, and $\mathbf{A}\mathbf{A}^{\#}=\mathbf{A}^{\#}\mathbf{A}$.

3. Differentiation of the stationary distribution. Throughout this section, we assume that $\mathbf{A}(t)=\mathbf{I}-\mathbf{P}(t)$ where $\mathbf{P}(t)$ is a matrix which is row stochastic and irreducible for each t in some interval (a,b). Furthermore, we will assume that each entry $p_{ij}(t)$ of $\mathbf{P}(t)$ is differentiable at each t in (a,b). It is important to note at the outset that, in general, the null vectors of a differentiable matrix need not be differentiable. However, for our special situation, normalizing a null vector of $\mathbf{A}(t)$ so as to produce the stationary distribution vector $\boldsymbol{\pi}(t)$ always results in a differentiable vector.

THEOREM 3.1. *If $\mathbf{A}(t)=\mathbf{I}-\mathbf{P}(t)$ where $\mathbf{P}(t)$ is row stochastic, irreducible, and differentiable on (a,b), then each component of the unique stationary distribution $\boldsymbol{\pi}(t)$ of $\mathbf{P}(t)$ satisfying $\boldsymbol{\pi}(t)=\boldsymbol{\pi}(t)\mathbf{P}(t)$, $\sum\pi_i(t)=1,u$ $\pi_i(t)>0$, is differentiable on (a,b).*

Proof. If $\mathbf{D}_i(t)$ denotes the ith principal minor of $\mathbf{A}(t)$ obtained by deleting the ith row and ith column of $\mathbf{A}(t)$, then for each t in (a,b), $\mathbf{D}_i(t)>0$ and $\boldsymbol{\pi}(t)$ is given by

$$\boldsymbol{\pi}(t)=\frac{1}{\sum\mathbf{D}_i(t)}[\mathbf{D}_1(t),\mathbf{D}_2(t),\cdots,\mathbf{D}_m(t)]. \tag{3.1}$$

This formula for $\boldsymbol{\pi}(t)$ is a simple consequence of the fact that $(\operatorname{adj}\mathbf{A})\mathbf{A}=\mathbf{A}(\operatorname{adj}\mathbf{A})=0$ and $\{\mathbf{e}\}$ is a basis for $N(\mathbf{A})$. Because the entries of $\mathbf{A}(t)$ are differentiable, each $\mathbf{D}_i(t)$ must be differential and hence each component of $\boldsymbol{\pi}(t)$ must be differentiable at each t in (a, b). □

In the sequel, we will omit writing the argument t (e.g., instead of writing $\boldsymbol{\pi}(t)$, simply write $\boldsymbol{\pi}$) and we sometimes will also use the dot notation $(\dot{\ })$ to indicate differentiation with respect to t (e.g., write $\dot{\boldsymbol{\pi}}$ instead of $d\boldsymbol{\pi}(t)/dt$).

THEOREM 3.2. *If $\mathbf{P}=\mathbf{P}(t)$ is row stochastic, irreducible, and differentiable for t in (a, b), then the derivative of the stationary distribution associated with $\mathbf{P}$ is given by*

$$\dot{\boldsymbol{\pi}}=\boldsymbol{\pi}\dot{\mathbf{P}}\mathbf{A}^{\#} \tag{3.2}$$

where $\mathbf{A}^{\#}$ denotes the group inverse of $\mathbf{A}=\mathbf{I}-\mathbf{P}$ as described in the previous section.

Proof. Use the elementary product rule for differentiation on $\boldsymbol{\pi}\mathbf{A}=\mathbf{0}$ to produce $\dot{\boldsymbol{\pi}}\mathbf{A}+\boldsymbol{\pi}\dot{\mathbf{A}}=\mathbf{0}$ or

$$\dot{\boldsymbol{\pi}}\mathbf{A}=\boldsymbol{\pi}\dot{\mathbf{P}}. \tag{3.3}$$

In general, if $\mathbf{Bx}=\mathbf{c}$ where $\mathbf{c}\in R(\mathbf{B})$ and $\dim N(\mathbf{B})=1$, then $\mathbf{x}$ must be given by $\mathbf{x}=\mathbf{B}^{\#}\mathbf{c}+\mathbf{n}$ for some $\mathbf{n}\in N(\mathbf{B})$. Apply this (using transposes) to (3.3) in order to obtain

$$\dot{\boldsymbol{\pi}}=\boldsymbol{\pi}\dot{\mathbf{P}}\mathbf{A}^{\#}+\alpha\boldsymbol{\pi} \quad \text{for some } \alpha. \tag{3.4}$$

To determine α, differentiate $\boldsymbol{\pi}\mathbf{e}=1$ to obtain $\dot{\boldsymbol{\pi}}\mathbf{e}=0$ and apply this together with the fact that $\mathbf{e}\in N(\mathbf{A})=N(\mathbf{A}^{\#})$ to (3.4). Thus

$$0=\dot{\boldsymbol{\pi}}\mathbf{e}=\boldsymbol{\pi}\dot{\mathbf{P}}\mathbf{A}^{\#}\mathbf{e}+\alpha\boldsymbol{\pi}\mathbf{e}=\alpha$$

and therefore

$$\dot{\boldsymbol{\pi}}=\boldsymbol{\pi}\dot{\mathbf{P}}\mathbf{A}^{\#}. \qquad □$$

By multiplying (3.2) on the right by the ith unit column $\mathbf{e}_i$, we may extract the expression for the derivative of the ith stationary probability $\boldsymbol{\pi}_i$.

COROLLARY. *The derivative of the ith stationary probability is given by,*

$$\dot{\boldsymbol{\pi}}_i=\boldsymbol{\pi}\dot{\mathbf{P}}\mathbf{A}_i^{\#} \tag{3.5}$$

where $\mathbf{A}_i^{\#}$ is the ith column of $\mathbf{A}^{\#}$.

One of the most pleasing aspects of Theorem 3.2 and its corollary is the sheer simplicity. The simple structure of (3.2) and (3.5) make it absolutely clear how the stationary distribution changes as the transition probabilities change. It shows that $\mathbf{A}^{\#}$ acts as a "magnification factor". If, at a particular point t_0, the derivatives of the transition probabilities are all relatively small and the ith column of $\mathbf{A}^{\#}$ contains only relatively small entries, then the ith stationary probability $\boldsymbol{\pi}_i$ must have a relatively small derivative. Because the entries of $\mathbf{A}^{\#}=\mathbf{A}^{\#}(t)$ are continuous functions of t (Corollary 3.1 in Meyer [10]), it follows that at t_0, $\boldsymbol{\pi}_i$ cannot be extremely sensitive to small perturbations in the transition probabilities whenever the ith column of $\mathbf{A}^{\#}(t_0)$ has no entries of relatively large magnitude. On the other hand, if the ith column of $\mathbf{A}^{\#}(t_0)$ contains some entries of large magnitude, then small perturbations in $\mathbf{P}(t_0)$ can be greatly magnified so as to make $\boldsymbol{\pi}_i$ very sensitive near t_0.

More precisely, translate the discussion to the origin and write

$$\boldsymbol{\pi}_i(t)-\boldsymbol{\pi}_i(0)=\dot{\boldsymbol{\pi}}_i(0)t+O(t^2)$$

and

$$t\dot{\mathbf{P}}(0)=\mathbf{P}(t)-\mathbf{P}(0)+O(t^2).$$

Theorem 3.2 now produces the following perturbation formula.

$$\pi_i(t)-\pi_i(0)=\pi(0)[\mathbf{P}(t)-\mathbf{P}(0)]\mathbf{A}_i^{\#}(0)+O(t^2). \tag{3.6}$$

It is transparent from (3.6) that the entries of $\mathbf{A}^{\#}(0)$ are the fundamental quantities which govern the sensitivity of the stationary probabilities. Assuming that t is small enough so that higher order terms may be neglected, apply Hölder's inequality to (3.6) and obtain the inequality

$$|\pi_i(t)-\pi_i(0)|\leqq\|\pi(0)\|_p\|[\mathbf{P}(t)-\mathbf{P}(0)]\mathbf{A}_i^{\#}(0)\|_q \tag{3.7}$$

where $(1/p)+(1/q)=1$. For all of the Hölder norms it is the case that $\|\pi(t)\|_p\leqq 1$. Thus for every Hölder vector norm $\|\cdot\|_q$ and compatible matrix norm $\|\cdot\|_m$, it follows from (3.7) that

$$|\pi_i(t)-\pi_i(0)|\leqq\|\mathbf{P}(t)-\mathbf{P}(0)\|_m\|\mathbf{A}_i^{\#}(0)\|_q. \tag{3.8}$$

The observations made throughout this section motivate the following definition.

DEFINITION. For an ergodic Markov chain with transition matrix $\mathbf{P}$ and stationary distribution π, the *condition number for the ith stationary probability* π_i is defined to be the number

$$\text{Cond}_q\,(\pi_i)=\|\mathbf{A}_i^{\#}\|_q$$

where $\|\cdot\|_q$ is any Hölder vector norm and $\mathbf{A}_i^{\#}$ is the ith column of the group inverse of $\mathbf{A}=\mathbf{I}-\mathbf{P}$. For a matrix norm $\|\cdot\|_m$, the number

$$\text{Cond}_m\,(\pi)=\|\mathbf{A}^{\#}\|_m$$

is defined to be the *condition number for* π. This number will also be referred to as the *condition of the underlying Markov chain.*

4. Linear perturbations. A special case of the preceding analysis which is of particular interest is that in which the perturbations are linear functions. That is, for a fixed row stochastic irreducible matrix $\mathbf{P}_0$, let $\mathbf{F}$ be a constant matrix such that

$$\mathbf{P}(t)=\mathbf{P}_0+t\mathbf{F}$$

is row stochastic and irreducible on (a, b). As before, let $\mathbf{A}(t)=\mathbf{I}-\mathbf{P}(t)$ and $\mathbf{A}_0=\mathbf{A}(0)=\mathbf{I}-\mathbf{P}_0$. By making use of our earlier results, we can obtain a very simple and explicit formula for $\dot{\pi}$, the derivative of the stationary distribution associated with $\mathbf{P}(t)$.

THEOREM 4.1. *If* $\mathbf{P}(t)=\mathbf{P}_0+t\mathbf{F}$ *is row stochastic and irreducible on* $[0, \beta)$, *then the derivative of the stationary distribution* $\pi(t)$ *associated with* $\mathbf{P}(t)$ *is given by*

$$\dot{\pi}(t)=\pi(t)\mathbf{F}\mathbf{A}_0^{\#}[\mathbf{I}-t\mathbf{F}\mathbf{A}_0^{\#}]^{-1}\quad \textit{for } t \textit{ in } [0,\beta) \tag{4.1}$$

where $\mathbf{A}^{\#}$ *is the group inverse of* $\mathbf{A}_0=\mathbf{I}-\mathbf{P}_0$.

Proof. Using Theorem 3.2, we obtain

$$\dot{\pi}(t)=\pi(t)\mathbf{F}(\mathbf{I}-\mathbf{P}_0-t\mathbf{F})^{\#}=\pi(t)\mathbf{F}(\mathbf{A}_0-t\mathbf{F})^{\#}. \tag{4.2}$$

Let $\pi_0=\pi(0)$. From Meyer [10, Thm. 3.1], the matrix $(\mathbf{I}-t\mathbf{F}\mathbf{A}_0^{\#})$ is always nonsingular and the term $(\mathbf{A}_0-t\mathbf{F})^{\#}$ can be expanded as follows.

$$(\mathbf{A}_0-t\mathbf{F})^{\#}=\mathbf{A}_0^{\#}+t\mathbf{A}_0^{\#}\mathbf{F}\mathbf{A}_0^{\#}(\mathbf{I}-t\mathbf{F}\mathbf{A}_0^{\#})^{-1}-\mathbf{P}_0^{\infty}(\mathbf{I}-t\mathbf{F}\mathbf{A}_0^{\#})^{-1}\mathbf{A}_0^{\#}(\mathbf{I}-t\mathbf{F}\mathbf{A}_0^{\#})^{-1} \tag{4.3}$$

where $\mathbf{P}_0^{\infty}=\mathbf{e}\pi$ is as described in (2.3). Since

$$\mathbf{e}=\mathbf{P}(t)\mathbf{e}=\mathbf{P}_0\mathbf{e}+t\mathbf{F}\mathbf{e}=\mathbf{e}+t\mathbf{F}\mathbf{e}$$

must hold for all t in $[0, \beta)$, it follows that $\mathbf{Fe}=\mathbf{0}$ and hence

$$\mathbf{F}\mathbf{P}_0^\infty = \mathbf{F}\mathbf{e}\boldsymbol{\pi}_0 = \mathbf{0}.$$

Use this when substituting (4.3) into (4.2) to obtain

$$\dot{\boldsymbol{\pi}}(t) = \boldsymbol{\pi}(t)\mathbf{F}\mathbf{A}_0^\#[\mathbf{I} + t\mathbf{F}\mathbf{A}_0^\#(\mathbf{I} - t\mathbf{F}\mathbf{A}_0^\#)^{-1}]. \tag{4.4}$$

By making use of the identity

$$\mathbf{I} = (\mathbf{I} - t\mathbf{F}\mathbf{A}_0^\#)^{-1} - t\mathbf{F}\mathbf{A}_0^\#(\mathbf{I} - t\mathbf{F}\mathbf{A}_0^\#)^{-1},$$

(4.4) reduces to

$$\dot{\boldsymbol{\pi}}(t) = \boldsymbol{\pi}(t)\mathbf{F}\mathbf{A}_0^\#(\mathbf{I} - t\mathbf{F}\mathbf{A}_0^\#)^{-1},$$

which is the desired conclusion. □

There are at least two interesting features to this theorem. The first is to notice that at $t=0$, the behavior of $\dot{\boldsymbol{\pi}}(t)$ is governed strictly by $\mathbf{F}$, $\mathbf{A}_0^\#$, and $\boldsymbol{\pi}(0)$.

COROLLARY. *For $\mathbf{P}(t) = \mathbf{P}_0 + t\mathbf{F}$, the derivative $\dot{\boldsymbol{\pi}}(t)$ of the stationary distribution evaluated at $t=0$ is given by*

$$\dot{\boldsymbol{\pi}}(0) = \boldsymbol{\pi}(0)\mathbf{F}\mathbf{A}_0^\#. \tag{4.5}$$

The derivative $\dot{\pi}_i(t)$ of the ith stationary probability at $t=0$ is

$$\dot{\pi}_i(0) = \boldsymbol{\pi}(0)\mathbf{F}[\mathbf{A}_0^\#]_i \tag{4.6}$$

where $[\mathbf{A}_0^\#]_i$ is the ith column of $\mathbf{A}_0^\#$. For any Hölder vector norm and compatible matrix norm,

$$|\dot{\pi}_i(0)| \leqq \|\mathbf{F}[\mathbf{A}_0^\#]_i\| \leqq \|\mathbf{F}\|\ \|[\mathbf{A}_0^\#]_i\|. \tag{4.7}$$

Another important point to be made concerning Theorem 4.1 and its corollary is the fact that *neither t nor $\mathbf{F}$ is required to be "small"*. The formula for $\dot{\boldsymbol{\pi}}$ in (4.1) as well as those in (4.5)-(4.7) are global in the sense that they hold for all t and $\mathbf{F}$ for which $\mathbf{P}(t) = \mathbf{P}_0 + t\mathbf{F}$ represents an irreducible transition matrix. However, if either t or $\mathbf{F}$ is small enough in magnitude to insure that $\|t\mathbf{F}\mathbf{A}_0^\#\| < 1$ then for compatible vector and matrix norms such that $\|\mathbf{I}\| \leqq 1$

$$(\mathbf{I} - t\mathbf{F}\mathbf{A}_0^\#)^{-1} = \sum_{k=0}^{\infty} (t\mathbf{F}\mathbf{A}_0^\#)^k$$

so that taking norms in (4.1) produces the following corollary.

COROLLARY. *If $\|t\mathbf{F}\mathbf{A}_0^\#\| < 1$, then*

$$\|\dot{\boldsymbol{\pi}}\| \leqq \|\boldsymbol{\pi}\| \frac{\|\mathbf{F}\mathbf{A}_0^\#\|}{1 - t\|\mathbf{F}\mathbf{A}_0^\#\|}. \tag{4.8}$$

Furthermore, if $t\|\mathbf{F}\|\ \|A_0^\#\| < 1$, then

$$\frac{\|\dot{\boldsymbol{\pi}}\|}{\|\boldsymbol{\pi}\|} \leqq \frac{\dfrac{\|\mathbf{F}\|}{\|\mathbf{A}_0\|}\|\mathbf{A}_0\|\ \|\mathbf{A}_0^\#\|}{1 - t\dfrac{\|\mathbf{F}\|}{\|\mathbf{A}_0\|}\|\mathbf{A}_0\|\ \|\mathbf{A}_0^\#\|} = \frac{\dfrac{\|\mathbf{F}\|}{\|\mathbf{A}_0\|}\kappa(\mathbf{A}_0)}{1 - t\dfrac{\|\mathbf{F}\|}{\|\mathbf{A}_0\|}\kappa(\mathbf{A}_0)}. \tag{4.9}$$

The expression (4.9) is a continuous counterpart of the discrete formula given by Meyer in [10].

The results of §§ 3 and 4 make it clear that for a finite homogeneous ergodic Markov chain, the sensitivity of the stationary probabilities are directly governed by the entries of the $\mathbf{A}^{\#}$ matrix. There appears to be ample evidence to support the use of $\mathbf{A}^{\#}$ as the fundamental quantity in gauging the "condition of a finite Markov chain" and it seems apparent that any perturbation or sensitivity analysis of a finite Markov chain should revolve around the matrix $\mathbf{A}^{\#}$.

5. Utilizing a QR factorization. The utility of orthogonal triangularization is well documented in the vast literature on matrix computations. The purpose of this section is to demonstrate how one can use a **QR** factorization of $\mathbf{A}=\mathbf{I}-\mathbf{P}$ to not only compute the stationary distribution $\boldsymbol{\pi}$, but to also gain some insight into the relative sensitivity which $\boldsymbol{\pi}$ is expected to exhibit.

For every $n\times n$ irreducible row stochastic matrix $\mathbf{P}$, it is well known that $\mathbf{A}=\mathbf{I}-\mathbf{P}$ must have Rank $(\mathbf{A})=n-1$. Moreover, any subset of $n-1$ columns from $\mathbf{A}$ is linearly independent. There is "essentially" a unique **QR** factorization of **A**. The **R**-factor is uniquely determined by **A** and the **Q**-factor is unique up to the algebraic sign of the last column.

THEOREM 5.1. *If* $\mathbf{A}_{n\times n}$ *is as described in the previous sections and if* $\mathbf{A}=\mathbf{QR}$ *is a* **QR** *factorization of* **A**, *then* **R** *must have the form*

$$\mathbf{R}=\left[\begin{array}{c|c}\mathbf{U} & -\mathbf{Ue}\\ \hline \mathbf{0} & \mathbf{0}\end{array}\right] \tag{5.1}$$

where **U** *is a nonsingular upper triangular* $(n-1)\times(n-1)$ *matrix and* **e** *is the column of* 1*'s. The stationary distribution* $\boldsymbol{\pi}$ *can be recovered from the last column,* **q**, *of* **Q** *as*

$$\boldsymbol{\pi}=\frac{\mathbf{q}^T}{\sum_{j=1}^n \mathbf{q}_j}. \tag{5.2}$$

Proof. To prove that **R** has the form (5.1), we need to show $r_{nn}=0$. Let **e** be the column of all 1's and use the fact that $\mathbf{0}=\mathbf{Ae}=\mathbf{QRe}$ to obtain

$$\mathbf{Re}=\mathbf{0}.$$

This together with the fact that **R** is upper triangular guarantees that $r_{nn}=0$. The fact that **U** is nonsingular now follows by noting that

$$\text{Rank}\,(\mathbf{U})=\text{Rank}\,(\mathbf{R})=\text{Rank}\,(\mathbf{QR})=\text{Rank}\,(\mathbf{A})=n-1.$$

To see that the stationary probabilities can be obtained from the last column of **Q**, recall that $\mathbf{A}=\mathbf{I}-\mathbf{P}$ where **P** is a nonnegative irreducible matrix with spectral radius 1. One consequence of the Perron-Frobenius theorem is that if $\mathbf{x}^T\mathbf{A}=\mathbf{0}$, then $\mathbf{x}^T>\mathbf{0}$ or $\mathbf{x}^T<0$. Moreover, the system

$$\mathbf{x}^T\mathbf{A}=\mathbf{0},\quad \mathbf{x}^T>0,\quad \|\mathbf{x}^T\|_1=1 \tag{5.3}$$

possesses a unique solution for $\mathbf{x}^T$. Since the last row of $\mathbf{R}=\mathbf{Q}^T\mathbf{A}$ is zero, it is clear that

$$\mathbf{0}=\mathbf{q}^T\mathbf{A}$$

where **q** is the last column of **Q** and hence $\mathbf{q}^T>\mathbf{0}$ or $\mathbf{q}^T<\mathbf{0}$. Thus

$$\mathbf{q}^T\Big/\sum_{j=1}^n \mathbf{q}_j$$

satisfies (5.3). Since the stationary distribution also satisfies (5.3), it must be the case that

$$\boldsymbol{\pi}=\frac{\mathbf{q}^T}{\sum_{j=1}^n \mathbf{q}_j}. \qquad \square$$

If the last column of the **Q**-factor produces the stationary distribution $\boldsymbol{\pi}$, of what relevance, if any, is the information in the **R**-factor? We now address this question and demonstrate how to use the **R**-factor to gauge the inherent sensitivity which $\boldsymbol{\pi}$ can exhibit to small perturbations in **P**.

Recall from the previous sections that the sensitivity of the ith stationary probability π_i is directly governed by the magnitude of the entries in the ith column of $\mathbf{A}^{\#}$. The goal is to estimate $\|\mathbf{A}^{\#}\|_2$ using only the information in the **R**-factor.

Unfortunately, group inversion is somewhat different than other familiar inversion processes in that $(\mathbf{QR})^{\#}\neq\mathbf{R}^{\#}\mathbf{Q}^T$. Moreover, there is no known way to directly unravel the term $(\mathbf{QR})^{\#}$ in terms of **Q** and **R**. However, the special structure of the matrix $\mathbf{A}=\mathbf{I}-\mathbf{P}$ does provide a special factorization for $\mathbf{A}^{\#}$ which is quite useful.

THEOREM 5.2. *Let* $\mathbf{A}=\mathbf{I}-\mathbf{P}$ *be as described in the previous sections and let* $\mathbf{A}=\mathbf{QR}$ *be a* **QR**-*factorization for* **A** *in which*

$$R=\left[\begin{array}{c|c}\mathbf{U} & -\mathbf{Ue}\\ \hline \mathbf{0} & \mathbf{0}\end{array}\right]$$

is described in Theorem 5.1. *If* $\boldsymbol{\pi}$ *is the stationary distribution for* **P**, *then* $\mathbf{A}^{\#}$ *is given by*

$$\mathbf{A}^{\#}=(\mathbf{I}-\mathbf{e}\boldsymbol{\pi})\left[\begin{array}{c|c}\mathbf{U}^{-1} & \mathbf{0}\\ \hline \mathbf{0} & \mathbf{0}\end{array}\right]\mathbf{Q}^T(\mathbf{I}-\mathbf{e}\boldsymbol{\pi})$$

where **e** *is a column of* 1'*s.*

Proof. If $\mathbf{A}^-$ denotes the matrix

$$\mathbf{A}^-=\left[\begin{array}{c|c}\mathbf{U}^{-1} & \mathbf{0}\\ \hline \mathbf{0} & \mathbf{0}\end{array}\right]\mathbf{Q}^T,$$

then it is straightforward to verify that

$$\mathbf{AA}^-\mathbf{A}=\mathbf{A}. \tag{5.4}$$

From (2.3) it follows that

$$(\mathbf{I}-\mathbf{e}\boldsymbol{\pi})=\mathbf{AA}^{\#}.$$

Using the relationships in (2.4) together with (5.4) produces

$$\begin{aligned}\mathbf{A}^{\#}&=\mathbf{A}^{\#}\mathbf{AA}^{\#}=\mathbf{A}^{\#}(\mathbf{AA}^-\mathbf{A})\mathbf{A}^{\#}=(\mathbf{AA}^{\#})\mathbf{A}^-(\mathbf{AA}^{\#})\\ &=(\mathbf{I}-\mathbf{e}\boldsymbol{\pi})\left[\begin{array}{c|c}\mathbf{U}^{-1} & \mathbf{0}\\ \hline \mathbf{0} & \mathbf{0}\end{array}\right]\mathbf{Q}^T(\mathbf{I}-\mathbf{e}\boldsymbol{\pi}).\end{aligned} \qquad \square$$

The factorization of Theorem 5.2 easily produces an upper bound for $\|\mathbf{A}^{\#}\|_2$ which depends solely on the **R**-factor.

THEOREM 5.3. *Let* $\mathbf{A}=\mathbf{I}-\mathbf{P}$ *where* **P** *is an* $(n\times n)$ *irreducible stochastic matrix and let* $\mathbf{A}=\mathbf{QR}$ *be a* **QR**-*factorization for* **A**. *If* **U** *is the* $(n-1)\times(n-1)$ *leading principal submatrix of* **R**, *then*

$$\|\mathbf{A}^{\#}\|_2\leqq 2(n-1)\|\mathbf{U}^{-1}\|_2.$$

Proof. Take norms on the factorization of Theorem 5.2 to obtain

$$\|\mathbf{A}^{\#}\|_2 \leqq \|\mathbf{I}-\mathbf{e}\boldsymbol{\pi}\|_2^2\|\mathbf{U}^{-1}\|_2.$$

It is known that for all square matrices **B** with real entries

$$\|\mathbf{B}\|_2^2 \leqq \|\mathbf{B}\|_F^2 = \text{Trace } (\mathbf{B}^T\mathbf{B}).$$

Thus

$$\|\mathbf{I}-\mathbf{e}\boldsymbol{\pi}\|_2^2 \leqq \text{Trace } [\mathbf{I}-\mathbf{e}\boldsymbol{\pi}-\boldsymbol{\pi}^T\mathbf{e}^T+\boldsymbol{\pi}^T\mathbf{e}^T\mathbf{e}\boldsymbol{\pi}] = n-2+n\|\boldsymbol{\pi}\|_2^2.$$

However,

$$\|\boldsymbol{\pi}\|_2 \leqq \|\boldsymbol{\pi}\|_1 = 1$$

so that

$$\|\mathbf{I}-\mathbf{e}\boldsymbol{\pi}\|_2^2 \leqq 2n-2 = 2(n-1)$$

and

$$\|\mathbf{A}^{\#}\|_2 \leqq 2(n-1)\|\mathbf{U}^{-1}\|_2. \qquad \square$$

Therefore, the condition number of the $(n-1)\times(n-1)$ leading principal submatrix **U** of **R** may be taken as an estimate (or as a measure in itself) of the condition of the underlying chain. Since **U** is upper triangular with positive diagonals, estimating $\text{Cond}_2(\mathbf{U})$ is not overly difficult (e.g., LINPACK methods can be used).

6. Conclusions.

For an ergodic chain with transition matrix **P**, a **QR** factorization of the matrix $\mathbf{A}=\mathbf{I}-\mathbf{P}$ yields complete information in the sense that both the stationary distribution $\boldsymbol{\pi}$ as well as a measure of the sensitivity of $\boldsymbol{\pi}$ to perturbations in **P** may be deduced.

1. *$\boldsymbol{\pi}$ is obtained by normalizing the last column of* **Q**.
2. *The sensitivity of the chain may be gauged by* Cond(**U**) *where* **U** *is the* $(n-1)\times(n-1)$ *leading principal submatrix of* **R**.

In general, it is well known that an upper triangular matrix may be ill-conditioned without possessing relatively small diagonal elements. However, for the special situation of an irreducible Markov chain, we have not been able to produce an example of an ergodic chain so that the factorization $\mathbf{A}=\mathbf{QR}$ yields an **R** in which $\|\mathbf{U}^{-1}\|$ is large but **U** has no small diagonals. In all of our computational experience, the sensitive chains always seem to force a diagonal entry of **U** to be relatively small. The more sensitive the chain, the smaller some diagonal of **U** becomes, so it seems. There is clearly need for further study.

REFERENCES

[1] A. BRANDT, *Algebra's multigrid theory*, preprint, 1983.

[2] S. L. CAMPBELL AND C. D. MEYER, *Generalized inverses of linear transformations*, Surveys and Reference Works in Mathematics, Pitman, London, 1979.

[3] J. CONLISK, *Comparative statics for Markov chains*, preprint, 1983.

[4] R. E. FUNDERLIC AND M. T. HEATH, *Linear compartmental analysis of ecosystems*, ORNL/IBP-71/4, Oak Ridge National Laboratory, Oak Ridge, TN, 1971.

[5] R. E. FUNDERLIC AND J. B. MANKIN, *Solution of homogeneous systems of equations arising from compartmental models*, SIAM J. Sci. Stat. Comp., 2 (1981), pp. 375–413.

[6] L. GALEONE, *the use of positive matrices for the analysis of the large time behavior of the numerical solution of reaction-diffusion systems*, Math. Comp., 41 (1983), pp. 461–474.

[7] L. KAUFMAN, *Matrix methods for queuing problems*, SIAM J. Sci. Stat. Comp., 4 (1984), pp. 525-552.

[8] J. G. KEMENY AND J. L. SNELL, *Finite Markov Chains*, Van Nostrand, New York, 1960.

[9] C. D. MEYER, *The role of the group generalized inverse in the theory of finite Markov chains*, SIAM Rev. 17 (1975), pp. 443-464.

[10] ———, *The condition of a finite Markov chain and perturbation bounds for the limiting probabilities*, this journal, 1 (1980), pp. 273-283,

[11] P. J. SCHWEITZER, *Perturbation theory and finite Markov chains*, J. Appl. Prob., 5 (1968), pp. 401-413.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

013

EXPANDERS AND DIFFUSERS*

MARSHALL W. BUCK†

Abstract. Expander graphs are ingredients for making concentrating, switching, and sorting networks, and are closely related to sparse, doubly-stochastic matrices called *diffusers*. The best explicit examples of diffusers are defined by means of the action of elements of the matrix group $SL(2, \mathbf{Z})$ on certain finite mathematical objects. Some corresponding, explicit expanders were introduced by Margulis. However, Gabber and Galil were the first to obtain good estimates for the expanders and produce from them a family of directed acyclic superconcentrators having density 271.8. In this paper we review various techniques for making expanders from diffusers. We also demonstrate asymptotic upper bounds on the strength of algebraically defined classes of degree k diffusers. Each upper bound is given as the norm of a diffusion operator on an infinite discrete group, and bounds for several examples are calculated. Numerical evidence is supplied in support of our conjecture that these bounds can be achieved by certain algebraically defined examples. The conjecture, if true, would lead to superconcentrators of density less than 58.

Key words. concentrator, directed graph, doubly-stochastic matrix, eigenvalue, expander, group, network, random walk, superconcentrator

AMS(MOS) subject classifications. 68E10(primary), 05C20, 15A51, 20G40, 60J10, 94C15, 65F50

1. Introduction.

DEFINITION. A finite, bipartite graph Γ of degree k with n input and n output vertices is called an (n, k, d) *expander* if for every input subset S of Γ, the set of output points connected to S, denoted $\Gamma(S)$, has size bounded below by the inequality:

$$|\Gamma(S)| \geqq \left(1 + d\left(1 - \frac{|S|}{n}\right)\right)|S|.$$

The quantity d is called the *expansion coefficient*.

Expanders are defined in Margulis [16] and in Gabber-Galil [11] where explicit constructions are given for arbitrarily large expanders of fixed degree k, all having a common expansion coefficient. For applications, it is important that k is small and that d is as large as possible. The best expanders presently known are not explicitly constructed but are shown to exist on the basis of counting arguments proving that "most" graphs of a given degree are "good" expanders [17]. We restrict ourselves in this paper to explicit constructions generalizing the work of Margulis, Gabber and Galil.

Expander graphs are used to construct bounded concentrator graphs.

DEFINITION. For $\theta < 1$, an (n, θ, k, α) *bounded concentrator* is a bipartite graph Γ with n inputs, θn outputs, at most kn edges, such that every input subset X with $|X| \leqq \alpha n$ maps to an output set $\Gamma(X)$ at least as large.

Any (n, θ, k, α) bounded concentrator is actually a *concentrator* in the following sense: for any input subset X with $|X| \leqq \alpha n$ there will be a set of $|X|$ disjoint edges connecting X to an equal number of outputs. Gabber and Galil use an (m, k, d) expander to make an (n, θ, k', α) bounded concentrator, where $n = m(p+1)/p$, $\theta = p/(p+1)$, $k' = (k+1)p/(p+1)$, and $\alpha = \frac{1}{2}$. Their construction is described in § 8. These bounded concentrator graphs are used to build superconcentrator networks.

DEFINITION. An (n, k) *superconcentrator* is a directed acyclic graph with n inputs and n outputs, having at most kn edges, such that for any choice of r inputs and r outputs there is a collection of r disjoint paths starting in the input set and ending in

* Received by the editors September 18, 1984, and in revised form April 19, 1985.

† Institute for Defense Analyses, Princeton, New Jersey 08540.

the output set. A *family of superconcentrators of density* k is a set of $(n, k+o(1))$ superconcentrators for $1 \leqq n < \infty$.

Pippenger [18] gives a recursive construction of superconcentrators in terms of bounded concentrators, which is generalized by Gabber and Galil as follows. Connect input i directly to output i for $1 \leqq i \leqq n$, using n edges. Given an input subset and an equinumerous output subset, use these direct lines, if possible, to connect inputs to outputs by paths of length 1. After that there remain fewer than $n/2$ unmatched inputs to be connected to a set of unmatched outputs. Use an $(n, \theta, k, \frac{1}{2})$ bounded concentrator to concentrate the unmatched inputs into a new "input space" of size θn. Similarly concentrate the unmatched outputs into an "output space" of size θn. Then use a size θn superconcentrator to match the unmatched. Start the recursive definition by using, say, a complete n^2 edge graph as superconcentrator when n is small. The result of this construction is summarized in Gabber-Galil's Lemma 8: "If we can construct for all n an $(n, \theta_n, k, \frac{1}{2})$ bounded concentrator, where $\theta_n = \theta + \varepsilon_n$, $0 \leqq \theta < 1$, $\varepsilon_n = o(1)$, and $k > 1$, then we can construct a family of linear superconcentrators of density $(2k+1)/(1-\theta)$."

Bounded concentrators are also used by Bassalygo and Pinsker [8] to make nonblocking switching networks using $O(N \log N)$ connections to handle N inputs, and more recently, by Ajtai, Komlós and Szemerédi [1], [2] to construct sorting networks using $O(N \log N)$ comparators to sort N inputs.

To a directed graph Γ of out-degree k (at most k directed arcs leaving each vertex) containing n vertices, associate the bipartite graph having input and output sets which are copies of Γ and having an edge from input x to output y if and only if there is a directed edge in Γ from x to y. Then Γ is said to be an (n, k, d) expander if the associated bipartite graph is.

DEFINITION. An (n, k, λ) *diffuser* is a doubly-stochastic n by n matrix M, having at most k nonzero entries in each row and in each column, and satisfying

$$\|Mx\|_2 \leqq \lambda \|x\|_2$$

for every n-vector $x = (x_1, \cdots, x_n)$ with $\sum_{i=1}^n x_i = 0$. If the matrix M is symmetric, then it is called a *symmetric* diffuser. The smallest value of λ for which M is an (n, k, λ) diffuser is called the *diffusion coefficient* of M. The minimum value for k is called the *degree* of M.

For the matrix M there is a first-order Markov process on n states, such that whenever $x = (x_1, \cdots, x_n)$ is a probability distribution on the states at time t, the distribution at time $t+1$ will be given by Mx. Since M is stochastic, the vector Mx will have the same sum as x itself. Consequently, the space of vectors of sum 0 is an invariant subspace for M. Furthermore, M is doubly-stochastic, so it leaves constant vectors fixed: i.e., $M1 = 1$. Let the linear operator M_0 denote the restriction of the operator M to the space of vectors with sum 0, furnished with the Euclidean norm. The diffusion coefficient of M is just the operator norm of M_0, written $\|M_0\|$, and tells us how fast an initial probability distribution converges in L^2 norm to the stationary, uniform distribution.

Given an (n, k, λ) diffuser M, we can define the *skeleton* of M to be the expander of order n, out-degree k, and in-degree k, containing directed arcs for precisely those transitions to which M assigns positive probability; i.e., there is an arc connecting i to j if and only if $m_{ji} > 0$. From the diffusion coefficient of M we can obtain lower bounds on the expansion coefficient of its skeleton graph. Following is the basic result of this type, showing that expanders are obtained from diffusers. It is similar to a result of Tanner [19].

PROPOSITION 2.1. *The skeleton graph of an (n, k, λ) diffuser is an $(n, k, 1-\lambda^2)$ expander.*

In some cases it is possible to show that the expansion coefficient d is somewhat larger than that predicted by the above proposition.

PROPOSITION 2.3. *Suppose that all the entries of an (n, k, λ) diffuser M are multiples of $1/k$. Then the skeleton graph of M is an (n, k, d) expander for*

$$d = \min\left\{\frac{1}{k-1}, \frac{k}{k-1}(1-\lambda^2)\right\}.$$

Our best expanders are obtained by using the above propositions. However, Gabber and Galil obtain expansion results from "symmetrized" diffusers. Suppose that $\sigma_1, \cdots, \sigma_k$ are k permutations of the set Γ. Let σ_0 denote the identity permutation. Define the unitary operator U_σ on the space of L^2 functions by

$$U_\sigma f(z) = f(\sigma^{-1}(z)).$$

Define the real symmetric averaging operator M by its effect on functions on the set Γ:

$$Mf = \frac{1}{2k}\sum_{i=1}^{k} U_{\sigma_i} f + \frac{1}{2k}\sum_{i=1}^{k} U_{\sigma_i^{-1}} f.$$

The matrix of M is an $(n, 2k, \lambda)$ diffuser, where $\lambda = \|M_0\|$. Meanwhile, define the directed graph Γ with out-degree $k+1$ by connecting each point x to the $k+1$ points $\sigma_0(x), \cdots, \sigma_k(x)$. Then Proposition 2.4 states that the directed graph Γ has expansion coefficient at least $d = 1-\lambda$.

The *extended skeleton* of a diffuser is obtained from the skeleton graph by adding loops at each vertex, possibly increasing the out-degree and in-degree. Alon and Milman [6] prove an isoperimetric inequality for graphs, and use it to obtain expansion results for the extended skeletons of symmetric diffusers. Their result appears here as Proposition 2.5.

In this paper we consider expander-diffuser families arising from the action of selected elements of the matrix group $G = SL(2, \mathbf{Z})$ on various natural objects, like the two families of finite groups $PSL(2, \mathbf{Z}/n\mathbf{Z})$ and $(\mathbf{Z}/n\mathbf{Z}) \times (\mathbf{Z}/n\mathbf{Z})$, and the family of finite sets $\mathbf{P}^1(\mathbf{F}_p)$ (the 1-dimensional projective spaces over finite fields with prime numbers of elements). Natural actions on the infinite lattice $\mathbf{Z} \times \mathbf{Z}$ and on the 2-torus $T^2 = (\mathbf{R}/\mathbf{Z}) \times (\mathbf{R}/\mathbf{Z})$ will also play a role in the proofs of diffusiveness in the finite cases.

The only examples of this type for which explicit diffusion coefficients have been obtained are those acting on the discrete torus $(\mathbf{Z}/n\mathbf{Z}) \times (\mathbf{Z}/n\mathbf{Z})$. The results are based on the method of Gabber and Galil [11], which produces "expanders" on the continuous torus T^2 from corresponding (Fourier dual) diffusers on the infinite discrete group $\mathbf{Z} \times \mathbf{Z}$. Their method then goes from T^2 to $(\mathbf{Z}/n\mathbf{Z}) \times (\mathbf{Z}/n\mathbf{Z})$, keeping d the same but increasing k; each of the transition possibilities is replaced by several variants, by composing with different translations of the discrete torus. These results are summarized in § 7.

We exhibit other families of diffusers by choosing sets of generators in $SL(2, \mathbf{Z})$, letting them act on the finite spaces $P^1(\mathbf{F}_p)$. We conjecture that the diffusion coefficients of these finite diffusers are determined approximately by the norm of the "covering" diffusion on the appropriate infinite subgroup of $SL(2, \mathbf{Z})$, and we present numerical evidence in § 9 supporting the conjecture. The diffusers will be best, for a given number of generators, if the generators freely generate the subgroup. For even degree k we need $k/2$ free generators, but for $k = 2m+1$ we need $m+1$ generators, one of order 2. Assuming the truth of the conjecture in the case $k = 4$, there will be explicit superconcentrators using only 78 edges per input. Thus we would approach the density 36 achieved by Bassalygo's inexplicit superconcentrators [7]. An optimal family of degree 7 diffusers, having diffusion coefficient at most $2\sqrt{6}/7$, by the conjecture, would produce superconcentrators of density 57.1587. (Here we would use the Alon-Milman inequality embodied in Proposition 2.5.)

For concreteness, we show how to make diffusers with $k=4$ from two elements that freely generate a subgroup of G. Define the elements A, B by

$$A=\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix},\qquad B=\begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix}.$$

The elements $\alpha=A^2$ and $\beta=B^2$ freely generate the subgroup of G defined by

$$\left\{\begin{pmatrix}a & b\\ c & d\end{pmatrix} \text{ such that } \begin{matrix}a\equiv 1 \bmod 4\\ b\equiv 0 \bmod 2\\ c\equiv 0 \bmod 2\\ d\equiv 1 \bmod 4\end{matrix}\right\}.$$

If G acts on a set S, we can define a diffuser corresponding to the first-order Markov process on the state space S, making the transitions $s\rightarrow\alpha\cdot s$, $s\rightarrow\alpha^{-1}\cdot s$, $s\rightarrow\beta\cdot s$, and $s\rightarrow\beta^{-1}\cdot s$ equally probable. Picking the G-set to be $\mathbf{P}^1(\mathbf{F}_p)$, where each matrix acts as a linear fractional transformation $z\rightarrow(az+b)/(cz+d)$, we obtain the four permutations:

$$z+2,\quad z-2,\quad \frac{1}{z^{-1}+2},\quad \frac{1}{z^{-1}-2}.$$

The conjecture says that the above diffuser is, for most primes p, a $(p, 4, \sqrt{3}/2)$ diffuser, where $\sqrt{3}/2$ is the norm of the following diffusion operator based on the subgroup of G. In the free subgroup on two generators α and β, we connect by arcs those pairs of group elements which are left multiples via one of the elements in $\{\alpha, \beta, \alpha^{-1}, \beta^{-1}\}$. We obtain thereby an infinite graph which is a tree of degree 4. In this tree consider the random walk having probability 1/4 of stepping in each of these four directions. Associate to the random walk the corresponding diffusion operator M acting on all L^2 functions on the group. The operator norm of M turns out to be $\sqrt{3}/2$. (In general, as shown by Kesten [13], the symmetric diffusion operator on an infinite tree of uniform degree k has operator norm $\|M\|=2\sqrt{k-1}/k$.) Only six of the primes less than 700 lead to diffusers with diffusion coefficient greater than $\sqrt{3}/2$.

Just how good an algebraically defined family of diffusers can be is determined by the norm of the covering diffusion operator on the infinite group, and that in turn is influenced by the group structure. In particular, for the class of *amenable* groups (see § 4) the diffusion norm is always 1, so these groups produce asymptotically poor diffuser families. An example of an amenable group is the "$ax+b$"-group H consisting of all transformations of the form $x\rightarrow ax+b$ where a and b are rational numbers. Complementing this result for diffusers are the results of Klawe [15] which show that H cannot be used as the basis of a good infinite family of expanders.

We show in § 6 that the operator norm of the covering diffusion operator on the infinite group gives an asymptotic lower bound on diffusion coefficients for an algebraically defined family of finite, symmetric diffusers. Let M be a symmetric, finitely generated, right invariant diffusion operator on G. If G acts on a finite set Γ, then M transfers to a diffuser M_Γ on Γ having diffusion coefficient λ_Γ. Proposition 6.1 states that there is a sequence of real numbers $b_1, b_2, \cdots$ converging to 0, such that for every finite G-set Γ we have $\lambda_\Gamma \geqq \|M\|-b_{|\Gamma|}$.

In § 5 we calculate $\|M\|$ for various algebraically defined, infinite, symmetric diffusers. (These are the covering diffusers.) We do this by first determining the *return generating function*

$$R(z)=\sum_{n=0}^{\infty} r_n z^n,$$

where r_n is the probability of the random walk being at the origin at time n, having started there at time 0. Then we use Proposition 3.1, which states that $\|M\|$ is the reciprocal of the radius of convergence of $R(z)$.

Among infinite diffusers of a fixed degree k, the homogeneous, isotropic tree diffusers give the smallest diffusion coefficient:

PROPOSITION 5.2. *If M is an infinite, doubly-stochastic, degree k diffuser, then* $\|M\| \geqq 2\sqrt{k-1}/k$.

The two families of expanders in Gabber-Galil [11], having $k=5$ and $k=7$, each come from infinite diffusers on $SL(2, \mathbf{Z})$ having degree 3. In § 7 we explain how these diffusers turn into degree 5 and 7 diffusers on $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$. Example 3 of § 5 shows that the first family of diffusers has diffusion coefficient $(1+\sqrt{8\sqrt{2}+13})/6\approx 0.988482$, poor for making expanders. Example 1 of § 5 shows that the second family of diffusers has diffusion coefficient at most $2\sqrt{2}/3$, leading, by means of a form of Proposition 2.3, to an expansion coefficient of at least 1/6 for the Gabber-Galil family of degree 7 expanders. Meanwhile, Gabber and Galil's worse estimate for the expansion coefficient, $d_0'=1-\sqrt{3}/2\approx 0.1340$, can be obtained from Proposition 2.4 applied to the $\{\alpha, \beta, \alpha^{-1}, \beta^{-1}\}$ diffuser, which has diffusion norm $\sqrt{3}/2$. Our better estimate of d for their $k=7$ expanders permits us to construct superconcentrators of density 222, an improvement over the density 262 announced by Chung [9].

Using the $\{\alpha, \beta, \alpha^{-1}, \beta^{-1}\}$ diffuser directly, we obtain Proposition 7.3, which provides a family of $(n^2, 12, 1/3)$ expanders on $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$. Instead of the four transitions $\{\alpha, \beta, \alpha^{-1}, \beta^{-1}\}$ which give on $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$

$$(x, y)\rightarrow(x\pm 2y, y),$$
$$(x, y)\rightarrow(x, y\pm 2x),$$

the proven expander has 12 transitions defined by

$$(x, y)\rightarrow(x\pm(2y+a), y),$$
$$(x, y)\rightarrow(x, y\pm(2x+a))$$

where $a=0, 1, 2$. The purpose of the perturbations by 0, 1, and 2 is to be able to reduce the expansion property of $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$ to that for the continuous torus T^2 of which it can be considered a part. These "more symmetric" expanders lead to superconcentrators of density 190. Alon and Milman [5] have used nearly identical expanders of degree 13 to make superconcentrators of density 157.4.

In § 9 we provide numerical evidence, for primes less than 700, supporting our conjecture for diffusers acting on the finite projective lines $P^1(\mathbf{F}_p)$.

2. Diffusers are expanders. Let the diffuser M be the transpose of the transition matrix for a Markov process on a finite set Γ. Recall that the *skeleton* of M is the directed graph which has an edge connecting two points whenever there is a nonzero probability of transition from one point to the other. We can also use M to denote the diffusion operator on functions on Γ. Define the operator M_0 obtained by restricting M to those functions on Γ having sum 0. Then the L^2 operator norm $\|M_0\|$ is the diffusion coefficient of M. For an (n, k, λ) diffuser we henceforth assume that λ is as small as possible: $\lambda = \|M_0\|$. We have the following result, allowing us to prove expansion properties for skeleton graphs of diffusers.

PROPOSITION 2.1. *The skeleton graph of an (n, k, λ) diffuser is an $(n, k, 1-\lambda^2)$ expander.*

Proof. Let Γ be the skeleton graph of the diffuser M. Suppose we denote the uniform measure on Γ by μ, normalized so that $\mu(\Gamma) = 1$. Let E be a subset of Γ and let $|\Gamma| = n$. Then $\mu(E) = |E|/n$, so we must prove that $\mu(\Gamma(E)) \geqq (1 + d(1 - \mu(E)))\mu(E)$. Denoting the complement of E by E^c, we note that $\mu(E^c) = 1 - \mu(E)$ so that we must show that $\mu(\Gamma(E)) \geqq (1 + d\mu(E^c))\mu(E)$. Define the function f_E by

$$f_E(x) = \begin{cases} \mu(E)^{-1} & \text{for } x \in E, \\ 0 & \text{for } x \in E^c. \end{cases}$$

Taking the L^2 norm with respect to the measure μ, we have $\|f_E\|_2 = 1/\sqrt{\mu(E)}$. The function $f_{\Gamma(E)}$ has the least norm of all functions of integral 1 having nonzero support contained in $\Gamma(E)$. The function $M(f_E)$ is supported in the set $\Gamma(E)$, so we have

$$\begin{aligned} \mu(\Gamma(E))^{-1} &= \|f_{\Gamma(E)}\|^2 \leqq \|M(f_E)\|^2 \\ &\leqq \|M(1 + (f_E - 1))\|^2 = 1 + \|M(f_E - 1)\|^2 \leqq 1 + \lambda^2 \|f_E - 1\|^2. \end{aligned}$$

However

$$\|f_E - 1\|^2 = \frac{\mu(E^c)}{\mu(E)}$$

so we have

$$\begin{aligned} \mu(\Gamma(E))^{-1} \leqq 1 + \lambda^2 \frac{\mu(E^c)}{\mu(E)} &\leqq (\mu(E) + \lambda^2 \mu(E^c))\mu(E)^{-1} \\ &\leqq (1 - (1 - \lambda^2)\mu(E^c))\mu(E)^{-1}. \end{aligned}$$

Thus we have

$$\begin{aligned} \mu(\Gamma(E)) &\geqq (1 - (1 - \lambda^2)\mu(E^c))^{-1}\mu(E) \\ &\geqq (1 + (1 - \lambda^2)\mu(E^c))\mu(E). \end{aligned}$$

Thus Γ has expansion coefficient at least $d = 1 - \lambda^2$. Q.E.D.

In case that Γ is infinite we have the more elementary result

$$\mu(\Gamma(E)) \geqq \|M\|^{-2}\mu(E),$$

where μ is now the counting measure.

Note that we have also proved the following slightly stronger result, which is similar to the connection between eigenvalues and expansion appearing in Tanner [19].

PROPOSITION 2.2. *The skeleton graph Γ of an (n, k, λ) diffuser M is an expander having the following expansion property*:

$$\mu(\Gamma(E)) \geqq (1 - (1 - \lambda^2)\mu(E^c))^{-1}\mu(E),$$

where μ is the uniform probability measure on Γ.

In some cases it is possible to show that the expansion coefficient d is somewhat larger than that predicted by the above Propositions.

PROPOSITION 2.3. *Suppose that all the entries of an (n, k, λ) diffuser M are multiples of $1/k$. Then the skeleton graph of M is an (n, k, d) expander for*

$$d=\min\left\{\frac{1}{k-1}, \frac{k}{k-1}(1-\lambda^2)\right\}.$$

Proof. Looking at the proof of Proposition 2.1, we see that the only inequality subject to improvement is the one asserting that the norm of $f_{\Gamma(E)}$ must not be greater than the norm of the function $M(f_E)$:

$$\mu(\Gamma(E))^{-1}=\|f_{\Gamma(E)}\|^2 \leqq \|M(f_E)\|^2.$$

But $M(f_E)$ is constrained to take values which are multiples of $1/k\mu(E)$. We may assume that $\mu(E)\neq 0$ and $\mu(\Gamma(E))/\mu(E)\leqq k/(k-1)$; otherwise, because $(1+1/(k-1))\geqq 1+(1/(k-1))\mu(E^c)$, the expansion rate would be at least $1/(k-1)$ and there would be nothing left for us to prove. Then the most uniform function supported on $\Gamma(E)$ subject to the value constraint would take only the two values $(k-1)/k\mu(E)$ and $1/\mu(E)$, taking the first value on a subset of measure $k(\mu(\Gamma(E))-\mu(E))$ and the second value on a set of measure $k\mu(E)-(k-1)\mu(\Gamma(E))$. Thus, one can compute the norm of this "most uniform" function and compare it to that of $M(f_E)$:

$$\|M(f_E)\|^2 \geqq \mu(E)^{-2}\left\{\left(2-\frac{1}{k}\right)\mu(E)-\left(1-\frac{1}{k}\right)\mu(\Gamma(E))\right\}.$$

Combining with the inequality

$$\|M(f_E)\|^2 \leqq 1+\lambda^2\frac{\mu(E^c)}{\mu(E)}$$

from before (in the proof of Proposition 2.1), we obtain

$$\mu(E)^2+\lambda^2\mu(E^c)\mu(E)\geqq \frac{2k-1}{k}\mu(E)-\frac{k-1}{k}\mu(\Gamma(E)).$$

We then rewrite this to get a lower bound on $\mu(\Gamma(E))$:

$$\mu(\Gamma(E))\geqq \frac{k}{k-1}\left\{\frac{2k-1}{k}\mu(E)-\mu(E)^2-\lambda^2\mu(E)\mu(E^c)\right\}$$

and

$$\frac{\mu(\Gamma(E))}{\mu(E)}\geqq \frac{2k-1}{k-1}-\frac{k}{k-1}(\mu(E)+\lambda^2\mu(E^c))\geqq 1+\frac{k}{k-1}(1-\lambda^2)\mu(E^c). \qquad \text{Q.E.D.}$$

Results from symmetric expanders can be used to deduce expansion coefficients for unsymmetric ones. This is done in Gabber-Galil [11]. Suppose that $\sigma_1, \cdots, \sigma_k$ are k permutations of the set Γ. Let σ_0 denote the identity permutation. Define the unitary operator U_σ on the space of L^2 functions by

$$U_\sigma f(z)=f(\sigma^{-1}(z)).$$

Define the real symmetric averaging operator M by its effect on functions on the set Γ:

$$Mf=\frac{1}{2k}\sum_{i=1}^{k} U_{\sigma_i}f+\frac{1}{2k}\sum_{i=1}^{k} U_{\sigma_i^{-1}}f.$$

Meanwhile, define the directed graph Γ with out-degree $k+1$ by connecting each point x to the $k+1$ points $\sigma_0(x), \cdots, \sigma_k(x)$.

Proposition 2.4 (Gabber-Galil). *The directed graph* Γ *has expansion coefficient at least* $d = 1 - \|M_0\|$.

Proof. For each subset E of Γ, define the function g_E by

$$g_E = \begin{cases} \mu(E^c) & \text{on } E, \\ -\mu(E) & \text{on } E^c. \end{cases}$$

Then we have $\|g_E\|^2 = \mu(E)\mu(E^c)$. If A and B are subsets, let $A - B$ denote $A \cap B^c$. We then have

$$\|g_E - U_\sigma(g_E)\|^2 = \mu(E \oplus \sigma(E)) = 2\mu(\sigma(E) - E).$$

Thus

$$\begin{aligned} \sum_{i=1}^{k} \mu(\sigma_i(E) - E) &= \frac{1}{2} \sum_{i=1}^{k} \|g_E - U_{\sigma_i} g_E\|^2 \\ &= k\|g_E\|^2 - \sum_{i=1}^{k} \langle g_E, U_{\sigma_i} g_E\rangle \\ &= k\mu(E)\mu(E^c) - k\langle g_E, Mg_E\rangle. \end{aligned}$$

Thus,

$$\frac{1}{k} \sum_{i=1}^{k} \mu(\sigma_i(E) - E) = \mu(E)\mu(E^c) - \langle g_E, Mg_E\rangle \geqq (1 - \|M_0\|)\mu(E)\mu(E^c).$$

Therefore for one of the i we must have:

$$\mu(\sigma_i(E) - E) \geqq (1 - \|M_0\|)\mu(E)\mu(E^c),$$

so

$$\begin{aligned} \mu\left(\bigcup_{i=0}^{k} \sigma_i(E)\right) &\geqq \mu(E) + \mu(\sigma_i(E) - E) \geqq \mu(E) + (1 - \|M_0\|)\mu(E)\mu(E^c) \\ &\geqq (1 + (1 - \|M_0\|)\mu(E^c))\mu(E). \end{aligned}$$

Q.E.D.

Another path from diffusers to expanders is taken by Alon and Milman [6]. Their isoperimetric inequality for graphs can be stated in our language as follows.

Proposition 2.5 (Alon-Milman). *Let* Γ *be the skeleton graph of an* (n, k, λ) *symmetric diffuser, and let* μ *denote the uniform probability measure on* Γ. *Then for every subset E we have*

$$\mu(\Gamma(E) - E) \geqq \sqrt{\beta^2 + 4(1 - \mu(E))\mu(E)} - \beta$$

where

$$\beta = \frac{4\mu(E) + (1 - \lambda)^{-1} - 1}{2}.$$

Alon and Milman make a degree $k + 1$ expander from a degree k diffuser by forming the extended skeleton graph. They use the above inequality to estimate the expansiveness of the resulting expander.

3. Norms of diffusion operators on infinite graphs. Suppose we have a transitive, symmetric random walk, on an infinite graph Γ, having diffusion operator M acting on $L^2(\Gamma)$. Pick one point x_0 to be the *origin*. Then the L^2 operator norm of M, written $\|M\|$, is equal to the reciprocal of the radius of convergence of the *return generating* *function*

$$R(z)=\sum_{n=0}^{\infty} r_n z^n,$$

where r_n is the probability of the random walk being at the origin at time n, having started there at time 0. In other words we have

PROPOSITION 3.1. $\|M\|=\overline{\lim}_{n\to\infty}\sqrt[n]{r_n}$.

Proof. First note that $r_n=\langle\delta, M^n\delta\rangle$, where δ is the function concentrated at the origin, taking the value 1 there. By the Cauchy-Schwarz inequality $r_n\leqq\|M\|^n$. Hence we have one of the directions that we need: $\|M\|\geqq\overline{\lim}_{n\to\infty}\sqrt[n]{r_n}$. We will be finished if we can exhibit a subsequence of $\sqrt[n]{r_n}$ converging to $\|M\|$. In fact we propose to show that the subsequence determined by $n=2, 4, 8, 16, \cdots, 2^k, \cdots$ works. Let f be any L^2 function of norm 1. Notice that M is Hermitian and

$$\|Mf\|^2=\langle Mf, Mf\rangle=\langle M^2f, f\rangle\leqq\|M^2f\|.$$

Hence we have

$$\|Mf\|\leqq\|M^2f\|^{1/2}\leqq\|M^4f\|^{1/4}\leqq\cdots.$$

(We should mention that the above inequality implies easily that $\|M^{2^k}\|=\|M\|^{2^k}$ and therefore $\|M^n\|=\|M\|^n$ for all n.) Letting $f=\delta$, we see that the sequence $(r_{2^k})^{1/2^k}$ is monotone increasing and bounded from above by $\|M\|$, so it converges to a limit μ. We must show that $\mu\geqq\|M\|$. Since $\|M^n\delta\|=\sqrt{r_{2n}}$ for all n, we have $\|M^{2^k}\delta\|\leqq\mu^{2^k}$ for all k.

Since the diffusion is transitive, we can show that the radius of convergence of the return generating function is independent of the choice of origin. Transitivity implies that for each point x there will be some time K at which the random walk starting at the origin will have nonzero probability of arriving at x; i.e., we will have $\langle\delta_x, M^K\delta\rangle=\beta>0$. Later we will need the inequality:

CLAIM. $\|M^{2^k}\delta_x\|\leqq\beta^{-1}\|M^{2^k}\delta\|$.

Proof of claim. We know that $\delta_x\leqq\beta^{-1}M^K\delta$. Since M has only nonnegative transitions, we can apply any power of M to each side of the inequality. In particular we have $M^{2^k}\delta_x\leqq\beta^{-1}M^{2^k}M^K\delta$, and thereby have the norm inequality

$$\|M^{2^k}\delta_x\|\leqq\beta^{-1}\|M\|^K\|M^{2^k}\delta\|\leqq\beta^{-1}\|M^{2^k}\delta\|\leqq\beta^{-1}\mu^{2^k}.$$

The claim is established.

Suppose that f is any function of norm 1 supported on a finite subset $X\subseteq\Gamma$. Then we wish to show that $\|Mf\|\leqq\mu$. Since any L^2 function can be approximated by one of finite support and we know that $\|M\|\leqq 1$, the bound on M will hold and the proposition will be proved. Recall that

$$\|Mf\|\leqq\|M^2f\|^{1/2}\leqq\|M^4f\|^{1/4}\leqq\cdots.$$

Choose $\beta>0$ small enough so that for every x in X there will be some K such that $\langle\delta_x, M^K\delta\rangle\geqq\beta$. But $f=\sum_{x\in X}c(x)\delta_x$, so

$$\|Mf\|\leqq(\|M^{2^k}f\|)^{2^{-k}}\leqq\left(\sum_{x\in X}|c(x)|\,\|M^{2^k}\delta_x\|\right)^{2^{-k}}\leqq\left(\beta^{-1}\sum_{x\in X}|c(x)|\right)^{2^{-k}}\mu$$

for all k. Thus $\|Mf\|\leqq\mu$ and $\|M\|\leqq\mu$. Q.E.D.

In fact we always have the slightly stronger result:

PROPOSITION 3.2. $\|M\| = \lim_{n\to\infty} \sqrt[n]{r_{2n}} = \lim_{n\to\infty} \|M^n\delta\|^{1/n}$.

Proof. In the preceding proof we showed that $\|M^n\delta\|^{1/n}$ monotonically increases to the limit $\|M\|$ for $n=2^k$. Notice that for every $\lambda < \|M\|$ we will have a $K(\lambda)$ such that $\|M^{2^k}\delta\| \geqq \lambda^{2^k}$ for all $k \geqq \log_2 K(\lambda)$. For any $n \geqq K(\lambda)$ choose a k such that $2^{k-1} \leqq n < 2^k$. Then

$$\|M^n\delta\| \geqq \|M^{2^k-n}\|^{-1}\|M^{2^k}\delta\| = \|M\|^{n-2^k}\|M^{2^k}\delta\| \geqq \|M\|^{n-2^k}\lambda^{2^k}$$

$$\geqq \|M\|^{n-2^k}\lambda^{2^k}\left(\frac{\lambda}{\|M\|}\right)^{2n-2^k} = \left(\frac{\lambda^2}{\|M\|}\right)^n.$$

As $\lambda \to \|M\|$ so does $\lambda^2/\|M\|$. Q.E.D.

For the random walks to be introduced in § 5 another important generating function is the *first-return generating function* $Q(z)$, which is the power series whose z^n coefficient, for $n>0$, is the probability that the random walk wanders back to the origin for the first time at time n. The generating functions R and Q are related by

$$R(z) = \frac{1}{1-Q(z)}.$$

We know that $|Q(z)|<1$ inside the unit circle, so the radius of convergence of R is at least 1. Of course, we already knew this because $\|M\|$ must always be less than or equal to 1.

4. Expanders and diffusers built from group actions. Let G be a group acting transitively on a set Γ, and let ϕ be a probability function on G with support written $\operatorname{supp}(\phi) = \{g \in G \mid \phi(g) > 0\}$. Then ϕ defines a doubly-stochastic Markov process on G (or on Γ) by allowing transitions $h \to gh$ (or $\gamma \to g\gamma$) with probability $\phi(g)$. The corresponding right invariant diffusion operator, M_G, acting on $L^2(G)$ (and the diffusion M_Γ on $L^2(\Gamma)$) are the left convolution operators with respect to ϕ:

$$M_G(f)(y) = \sum_{g\in G} \phi(g)f(g^{-1}y)$$

or

$$M_G f = \sum_{g\in G} \phi(g)T_g f$$

where T_g is the left translation operator defined by $T_gf(y) \equiv f(g^{-1}y)$. The T_g are unitary operators (in particular they have norm 1), so the operators M_G and M_Γ have norm at most 1 because they are averages of T_g operators. If $\phi(g) = \phi(g^{-1})$ for all g, then M_G and M_Γ will have symmetric matrices and correspond to symmetric diffusers. Furthermore, when $|\operatorname{supp}(\phi)| = k$ the matrices will have degree k, so the skeleton graphs will have out-degree and in-degree k.

Pick a point $x_0 \in \Gamma$ to be called the *origin*. Let K be the isotropy subgroup in G for the point x_0, so that Γ is isomorphic to the coset space G/K. We would like a condition on K for the diffusions on G and Γ to have the same norm (i.e., $\|M_G\| = \|M_\Gamma\|$).

A group K is said to be *amenable* if there is a right and left invariant mean on the vector space of bounded functions on K; namely, there exists a right and left invariant functional $\lambda \in L^\infty(K)^*$ with the properties $\|\lambda\| = 1$ and $\lambda(1) = 1$. A group is right amenable (or left amenable) if there is a right (or left) invariant mean. Basic facts about amenability can be found in Hewitt-Ross [12], pp. 230-245. In particular we list the following:

1. Abelian groups are amenable.
2. Finite groups are amenable.
3. A group is right amenable if and only if it is left amenable.
4. Every right (or left) amenable group is amenable.
5. A group is amenable if and only if every finitely generated subgroup of it is amenable.
6. If N is a normal subgroup of a group G, then G is amenable if and only if both N and G/N are amenable.

The following theorem is an improvement, due to Day [10], of a result of Kesten [14], which connects the concept of amenability to the norm of right invariant diffusion operators on a group.

THEOREM. *If K is an amenable group and ϕ is any probability function on K, then $\|M_\phi^{(p)}\| = 1$, where $M_\phi^{(p)}$ is the operator of left convolution by ϕ acting on $L^p(K)$ for $1 \leqq p \leqq \infty$. Conversely, if there is some probability function ϕ on a group K, with $\phi(g) = \phi(g^{-1})$ and $\operatorname{supp}(\phi)$ generating K, such that for some p with $1 < p < \infty$ we have $\|M_\phi^{(p)}\| = 1$, then K must be amenable.*

PROPOSITION 4.1. *If K is amenable and a subgroup of G, if $\Gamma = G/K$, and if ϕ is a symmetric probability function on G giving rise to the symmetric diffusion operators M_G and M_Γ on G and Γ respectively, then $\|M_G\| = \|M_\Gamma\|$. However, if K is not amenable, and $\operatorname{supp}(\phi)$ generates the group G (so that, in particular, M_G is transitive) then $\|M_G\| < \|M_\Gamma\| \leqq 1$.*

Proof. See Kesten [13]. (His proof is claimed only for the case in which K is a normal subgroup, but it generalizes easily.) We prove only the first assertion here. First note that $\|M_G\| \leqq \|M_\Gamma\|$ is always true for symmetric ϕ. In fact $M_G^n\delta$ collapses to $M_\Gamma^n\delta$, so $\|M_G^n\delta\|_2 \leqq \|M_\Gamma^n\delta\|_2$, and by Proposition 3.2 we have $\|M_G\| \leqq \|M_\Gamma\|$. Now we need to show that $\|M_G\| \geqq \|M_\Gamma\|$.

Let I_K be multiplication by the characteristic function of K. Notice that

$$\|M_G\|^{2n} = \|M_G^{2n}\| \geqq \|I_K M_G^{2n} I_K\|$$

and

$$I_K M_G^{2n} I_K = \langle \delta_{x_0}, M_\Gamma^{2n} \delta_{x_0} \rangle N_{2n}$$

where N_{2n} is a symmetric, right invariant diffusion on K. Since K is amenable, we have $\|N_{2n}\| = 1$. Hence we have

$$\|I_K M_G^{2n} I_K\| = |\langle \delta_{x_0}, M_\Gamma^{2n} \delta_{x_0} \rangle| = \|M_\Gamma^n \delta_{x_0}\|^2.$$

Thus, for all n

$$\|M_G\| \geqq \|M_\Gamma^n \delta_{x_0}\|^{1/n}$$

so

$$\|M_G\| \geqq \overline{\lim_{n\to\infty}} \|M_\Gamma^n \delta_{x_0}\|^{1/n} = \|M_\Gamma\|. \qquad \text{Q.E.D.}$$

A couple of easy applications of Proposition 4.1 are:

1. Let $G = SL(2, \mathbf{Z})$ and $K = \{I, -I\}$, the center of G. Then $G/K = PSL(2, \mathbf{Z})$ and the norm of a right invariant diffusion on G is the same as the norm of the induced diffusion on G/K.

2. Let G be the same, but let K be the Abelian subgroup consisting of matrices which have the form of I with an arbitrary integer in the upper right corner. We can

conclude that diffusions on $SL(2, \mathbf{Z})$ carry over, without reduction of norm, to diffusions on the set $\mathbf{Z}\times\mathbf{Z}-(0,0)$, given the natural action of G on it.

5. Some infinite diffusers and how well they work. Here we study random walks on infinite trees and on finitely generated subgroups of $PSL(2, \mathbf{Z})$. The case of infinite, homogeneous, degree k trees, with equal transition probabilities in all directions, is discussed as an example in [13] with the result that the operator norm is given by

$$\|M\|=\frac{2\sqrt{k-1}}{k}.$$

We can rederive that result here, making use of the first-return generating function. Define $T(z)$ to be the generating function giving the probability of first reaching the origin a after a number of steps from a starting point adjacent to a, say b. Then we have $Q(z)=zT(z)$. Notice that we can think of the value of $T(z)$, for $0\leqq z\leqq 1$, as the probability of ever reaching a, starting at b, given that we follow the random walk but also have the probability $1-z$ of dying on each step. Starting at a point c, instead, at distance m from a, the generating function is obtained by raising T to the power m. From the point b there are $k-1$ ways to go to a point twice removed from a, and only one way to go directly to a. Thus T must satisfy the identity:

$$T(z)=\frac{z}{k}+\frac{(k-1)z}{k}T^2(z).$$

Solving for T, we obtain

$$T(z)=\frac{k\pm\sqrt{k^2-4(k-1)z^2}}{2(k-1)z}.$$

Since $T(z)\leqq 1$ for $0\leqq z\leqq 1$ we must take only the minus sign above, for the other root will always be at least 1 when $k\geqq 2$. The numerator in the above expression has a zero of order 2 at the origin, considered as a function of a complex variable, so the radius of convergence of T is the distance to the branch point of the square root function. The branch point z_0 occurs at the zero of the discriminant, namely: $z_0=k/(2\sqrt{k-1})$. Thus, by Proposition 3.1, $\|M\|=2\sqrt{k-1}/k$.

The tree diffusion with even degree $k=2j$ is equivalent to that on a free group with j generators, where each generator and its inverse is used with probability $1/2j$. If $k=2j+1$, consider the group with $j+1$ generators, one of which has order 2. Of course, every finitely generated group is a factor group of a free group so, by Proposition 4.1, the diffusion coefficients will be no better than those for the free group.

Among all symmetric diffusions of degree k, the isotropic diffusion on the homogeneous, degree k tree has the smallest L^2 norm.

PROPOSITION 5.1. *If M is an infinite, symmetric, degree k diffuser, then* $\|M\|\geqq 2\sqrt{k-1}/k$.

Proof. Without loss of generality, assume that M is transitive. Cover M by a diffusion N based on the infinite, homogeneous, degree k tree T_k. This means that there is a surjective mapping $\pi: T_k\to\Gamma$ and the corresponding push-forward mapping π_* sending functions on T_k with finite support to functions on Γ, such that we have $\pi_*N=M\pi_*$.

A *sticky* vertex is one that has nonzero probability of transition to itself. We can construct N so it has no sticky vertices and allows transitions only to adjacent vertices in T_k. For each vertex in T_k that maps to a sticky vertex in Γ, there is an adjacent vertex that has the same image, and the sticky transition in Γ lifts to a symmetric transition between the pair of vertices in T_k.

We can apply Proposition 3.1 to show that $\|M\| \geqq \|N\|$. Thus, it suffices to prove our Proposition for the case in which M is based on T_k and has no sticky vertices.

We propose to show that $\|M\| \geqq 2\sqrt{k-1}/k$ by constructing a sequence of functions $f_1, f_2, f_3, \cdots$ such that

$$\varlimsup_{n\to\infty} \frac{\|Mf_n\|}{\|f_n\|} \geqq \frac{2\sqrt{k-1}}{k}.$$

Choose any vertex in T_k to be the root of the tree, and let $d(x)$ denote the distance of a vertex x from the root. Define f_n by

$$f_n(x) = \begin{cases} (k-1)^{-d(x)/2}, & d(x) \leqq n, \\ 0, & d(x) > n. \end{cases}$$

One computes easily that

$$\|f_n\|^2 = 1 + \frac{nk}{k-1}.$$

Let V_i denote the set of vertices at distance i from the root. Let $\bar{M}$ denote the isotropic diffusion, which assigns probability $1/k$ to every transition. (We have already shown that $\|\bar{M}\| = 2\sqrt{k-1}/k$.) Then

$$\sum_{x\in V_i} Mf_n(x) = \sum_{x\in V_i} \bar{M}f_n(x)$$

and $\bar{M}f_n$ is constant on V_i, so

$$\sum_{x\in V_i} |Mf_n(x)|^2 \geqq \sum_{x\in V_i} |\bar{M}f_n(x)|^2.$$

Summing over i, we obtain $\|Mf_n\| \geqq \|\bar{M}f_n\|$. But, an easy computation shows that

$$\|\bar{M}f_n\|^2 = \frac{1}{k-1} + \frac{4n-2}{k},$$

and

$$\lim_{n\to\infty} \frac{\|\bar{M}f_n\|^2}{\|f_n\|^2} = \frac{4(k-1)}{k^2}. \qquad \text{Q.E.D.}$$

PROPOSITION 5.2. *If M is an infinite, doubly-stochastic, degree k diffuser, then* $\|M\| \geqq 2\sqrt{k-1}/k$.

Proof. Suppose M has the skeleton graph Γ, which is an infinite, directed graph having out-degree and in-degree k. As described in the Introduction, let Γ_2 be the bipartite, undirected graph having input and output sets isomorphic to Γ. We can build a symmetric diffuser M_2 based on Γ_2 by using M and its transpose M^t. Then M_2 will have the (infinite) "matrix"

$$\begin{pmatrix} 0 & M^t \\ M & 0 \end{pmatrix}.$$

It is easy to see that $\|M_2\| = \|M\| = \|M^t\|$. But M_2 is an infinite, symmetric, degree k diffuser, so by Proposition 5.1 its norm satisfies the required inequality. Q.E.D.

The group $SL(2, \mathbf{Z})$ consists of all unimodular square matrices of size 2 with integer coefficients; i.e.,

$$SL(2, \mathbf{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ such that } a, b, c, d \in \mathbf{Z} \text{ and } ad - bc = 1 \right\}.$$

The group $PSL(2, \mathbf{Z})$ is defined to be $SL(2, \mathbf{Z})$ divided by its center, which contains only the two matrices I and $-I$. Thus we can think of an element of $PSL(2, \mathbf{Z})$ as a pair of unimodular matrices which are negatives of each other. The identity element e is the pair of matrices $\{I, -I\}$. $PSL(2, \mathbf{Z})$ has the structure of a group with two generators and two relations:

$$\{c, d \mid c^2 = d^3 = e\}.$$

The graph of the group consists of triangles, with looping provided by the group element d, connected at the corners by the involution (or "reflection") c. The element c is represented by the matrix

$$C = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

and the element d is represented by the matrix

$$D = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}.$$

Note that $C^2 = -I$ and $D^3 = -I$. The matrices A and B defined earlier are just $D^{-1}C$ and $-DC$ respectively.

Example 1. Define the two elements of $PSL(2, \mathbf{Z})$ by $\tau_1 = c$ and $\tau_2 = dcd$, generating a subgroup with the following structure:

$$\{\tau_1, \tau_2 \mid \tau_1^2 = e\}.$$

The graph of the subgroup is just the infinite, degree 3 tree, so the diffusion M that performs one of the three operations τ_1, τ_2, τ_2^{-1} with equal probability will have norm (as an operator acting on L^2) equal to $2\sqrt{2}/3$. Notice that $\tau_1 c = e = \sigma_0$, $\tau_2 c = dcdc = b^2$, $\tau_2^{-1}c = d^{-1}cd^{-1}c = a^2$, so the diffusion (no longer symmetric) defined by the three transitions $\{I, A^2, B^2\}$ will have the same norm. But it is precisely these transitions that are used as the basis of Gabber-Galil's $k = 7$ expanders. Thus, a generalization of our Proposition 2.3, used with $k = 3$, and combined with the analysis in § 7, shows that their example has expansion coefficient at least $d = 1/6$, thereby improving their estimate of $1 - \sqrt{3}/2$, which results from the use of Proposition 2.4 and Example 2, following.

Example 2. Define the four transitions by a^2, b^2 and their inverses. These freely generate the subgroup Γ_2 defined by

$$\Gamma_2 = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ such that } \begin{matrix} a \equiv 1 \bmod 2 \\ b \equiv 0 \bmod 2 \\ c \equiv 0 \bmod 2 \\ d \equiv 1 \bmod 2 \end{matrix} \right\}.$$

The graph of the group is the infinite, homogeneous, degree 4 tree. We have $\|M\| = \sqrt{3}/2$.

Example 3. Define three transitions by $\tau_1 = c$, $\tau_2 = d$, and $\tau_3 = \tau_2^{-1}$. The set $\{\tau_1, \tau_2, \tau_3\}$ will have the same expansiveness as its multiple by c: $\{\tau_1 c, \tau_2 c, \tau_3 c\} = \{e, a, b\}$. Thus we will be determining the expansiveness of Gabber-Galil's $k = 5$ example. Proposition 4.1 implies that $\|M\| \geqq 2\sqrt{2}/3$.

The graph of the group consists of triangles, navigated by the order 3 element d, linked together by edges that flip when c is applied.

The first-return generating function is

$$Q(z)=\frac{2z}{3}T(z)+\frac{z}{3}E(z)$$

where $T(z)$, for values of z between 0 and 1, can be thought of as the probability of ever getting from one to another of two vertices on the same triangle, when there is only a probability z of surviving each step, and where $E(z)$ is the probability of ever crossing a given edge, starting at one of its ends. Starting from a vertex, the probability of every crossing its bridge, $E(z)$, is just the probability of crossing immediately, $z/3$, plus the probability of crossing it after going to another point in the triangle, namely $(2z/3)T(z)E(z)$. Thus,

$$E(z)=\frac{z}{3}+\frac{2z}{3}T(z)E(z).$$

Similarly, we get

$$T(z)=\frac{z}{3}+\frac{z}{3}T(z)E(z)+\frac{z}{3}T(z).$$

We now change variables to $t=z/3$ for simplicity. Solving both equations for E, we get

$$E=\frac{t}{1-2tT}=\frac{(1-t)T-t}{tT}.$$

Clearing denominators, we get

$$2t(1-t)T^2-(1-t+t^2)T+t=0$$

leading to

$$T=\frac{1-t+t^2-\sqrt{D}}{4t(1-t)}$$

and

$$E=\frac{t}{1-2tT}=\frac{1-t-t^2-\sqrt{D}}{2t}$$

where D is the discriminant

$$D=(1-t+t^2)^2-8t^2(1-t)=t^4+6t^3-5t^2-2t+1.$$

Then,

$$Q=2tT+tE=\frac{1-t+t^2-\sqrt{D}}{2(1-t)}+\frac{1-t-t^2-\sqrt{D}}{2}.$$

The smallest root of D is

$$t_0=\frac{2}{1+\sqrt{8\sqrt{2}+13}}$$

and is the closest branch point for $\sqrt{D}$; thus, we have

$$\|M\|=\frac{1}{3t_0}=\frac{1+\sqrt{8\sqrt{2}+13}}{6}\approx 0.988482.$$

Example 4. Define the four transitions by $\{a, b, a^{-1}, b^{-1}\}$. If we multiply all four by c on the right, we obtain the elements $\{d^{-1}, d, cdc, cd^{-1}c\}$. Thus we consider the group generated by $\tau_1 = d$ and $\tau_2 = cdc$. It can be shown that the subgroup they generate is abstractly just

$$\{\tau_1, \tau_2 \mid \tau_1^3 = \tau_2^3 = e\}.$$

This subgroup of $PSL(2, \mathbf{Z})$ is that consisting of all elements that can be written as words in c, d using only an even number of c's. The graph consists of triangles touching each other at all of their vertices. As usual define the first-return generating function $Q(z) = zT(z)$, where $T(z)$ for values of z between 0 and 1 can be thought of as the probability of ever getting from one to the other of two adjacent points, when there is a probability z of surviving each step. If two points are at distance m apart in the graph, then the probability of ever making it to one, starting at the other, is just $T^m(z)$. From one point there is $z/4$ chance of success by going directly to the destination, there are two ways to go to points at distance 2 (from which the probability is $T^2(z)$ for getting to the destination), and there is one way to go to another adjacent point. Thus we must have the identity

$$T(z) = \frac{z}{4} + \frac{z}{4} T(z) + \frac{z}{2} T^2(z)$$

giving

$$2zT^2(z) + (z-4)T(z) + z = 0$$

with solution

$$T(z) = \frac{4 - z - \sqrt{(4-z)^2 - 8z^2}}{4z}.$$

The radius of convergence turns out to be given by the smallest root of the discriminant. The smallest root of the quadratic equation $7z^2 + 8z - 16 = 0$ is $4/(2\sqrt{2}+1)$. Thus we discover that $\|M\| = (2\sqrt{2}+1)/4$.

There is a way to arrive at the value of $\|M\|$ by relying upon the result for Example 1. At each point in the graph there are two ways to go away from the origin, one way to go towards it, and one way to stay at the same distance from it. The mixture $\frac{1}{4}I + \frac{3}{4}M_1$, where M_1 is the operator of Example 1, has the same return statistics to the origin. Hence,

$$\|M_4\| = \left\| \frac{1}{4} I + \frac{3}{4} M_1 \right\| = \frac{1}{4} + \frac{3}{4}\left(\frac{2\sqrt{2}}{3}\right) = \frac{2\sqrt{2}+1}{4}.$$

6. Asymptotic lower bounds for diffusion coefficients. We have been considering infinite families of finite diffusers constructed from actions of finitely generated infinite groups. Suppose that S is a finite set of generators for a finitely generated group G, and that S is closed under inversion. Consider the symmetric random walk on G in which we left multiply by elements of S chosen according to a probability distribution ϕ on S, invariant under inversion. The effect of the random walk on the vector space of all L^2 functions on G is determined by the diffusion operator M, which is invariant under right translation by group elements. Suppose that Γ is a finite set on which G acts. Then the random walk on G transfers to a random walk on Γ, producing an operator M_Γ on $L^2(\Gamma)$, and an operator $M_{\Gamma,0}$ on the space of functions with sum 0. In this section we use the counting measure on Γ, to be consistent with our use of the counting measure on the infinite group G.

We wish to show that the norm $\|M_{\Gamma,0}\|$ is bounded from below by $\|M\|$ in the limit as the number of elements in Γ goes to ∞. More precisely,

PROPOSITION 6.1. *Given a symmetric, finitely generated, right invariant, diffusion operator M on G, there exists a sequence of real numbers $b_1, b_2, \cdots$, converging to 0, such that for every finite G-set Γ we have*

$$\|M_{\Gamma,0}\| \geqq \|M\| - b_{|\Gamma|}.$$

Proof. Let S be the generating set for M. For $x \in \Gamma$, define $D_k(x)$ to be the set of all points in Γ reachable from x by means of at most k left multiplies by elements in the set S. As a crude estimate we will always have $|D_k(x)| \leqq m^{k+1}/(m-1)$ if S has m elements. Thus as $|\Gamma|$ goes to ∞, so does the diameter of Γ. Choose a function f such that every Γ has diameter at least $2f(|\Gamma|)+1$, and such that $f(n) \to \infty$ as $n \to \infty$.

Suppose that x and y are at distance $2k+1$; then the sets $D_k(x)$ and $D_k(y)$ are disjoint. Define the function $\eta = \delta_x - \delta_y$, obviously a function on Γ having sum 0. Because $M_\Gamma^k(\delta_x)$ is supported on $D_k(x)$, and $M_\Gamma^k(\delta_y)$ is supported on $D_k(y)$, we have

$$\|M_\Gamma^k \eta\|^2 = \|M_\Gamma^k \delta_x\|^2 + \|M_\Gamma^k \delta_y\|^2.$$

Now notice that $\|M_\Gamma^k \delta_x\| \geqq \|M^k \delta_e\|$ and $\|M_\Gamma^k \delta_y\| \geqq \|M^k \delta_e\|$ implies

$$2\|M_{\Gamma,0}\|^{2k} = \|M_{\Gamma,0}\|^{2k} \|\eta\|^2 \geqq \|M_{\Gamma,0}^k \eta\|^2 \geqq 2\|M^k \delta_e\|^2,$$

so we have

$$\|M_{\Gamma,0}\| \geqq \|M^k \delta_e\|^{1/k}.$$

But we know that the right-hand side of the above equation converges to $\|M\|$ as $k \to \infty$, by Proposition 3.2. Define $a_k = \|M\| - \|M^k \delta_e\|^{1/k}$, so that $\lim_{n \to \infty} a_n = 0$. Then we can choose $b_n = a_{f(n)}$. Q.E.D.

Example. Consider the "$ax+b$"-group H consisting of all transformations of the form $x \to ax+b$ where a and b are rational numbers. Let M be a finitely generated, symmetric, right invariant diffusion operator based on H. Let N be the normal subgroup of H consisting of all transformations of the form $x \to x+b$. Then N is isomorphic to the additive group of the rationals, and the quotient group H/N is isomorphic to the multiplicative group of the rationals. Both of these are Abelian groups and therefore are amenable. Hence H is amenable, and we must have $\|M\| = 1$. Thus finite diffusers built from M will degrade as the graphs grow larger. The corresponding assertion about the degradation of the expansion coefficient is the result proved by Klawe [15].

Alon [3] has shown that an undirected graph is a good expander if and only if a related symmetric diffuser is good; hence, it should be possible to prove results of the Klawe type for all amenable groups.

7. The construction of Gabber and Galil: going from the continuous torus to the discrete torus. The group $G = SL(2, \mathbf{Z})$ acts on the plane $\mathbf{R} \times \mathbf{R}$ in the natural way, the matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ sending the point (x, y) to the point $(ax+by, cx+dy)$. This action restricts to the lattice $\mathbf{Z} \times \mathbf{Z}$ in the plane, sending it onto itself. Since the 2-torus T^2 is just the plane modulo the lattice of integer points, the action on the plane factors to an action on the torus. Furthermore, the actions on the torus and on the lattice are related: the Fourier series for an L^2 function on the torus is an L^2 function on the lattice, and this Fourier transform intertwines the two representations of $SL(2, \mathbf{Z})$. Denote by τ the automorphism of G that sends a matrix to its inverse transpose: $\tau(\gamma) = (\gamma^*)^{-1}$. Let μ be Lebesgue measure on T^2, normalized so that $\mu(T^2) = 1$. Let Ff denote the Fourier transform of f; i.e.,

$$Ff(x)=\int_{T^2}\exp(-2\pi i\langle x, y\rangle)f(y)\,d\mu(y).$$

We then have:

PROPOSITION 7.1. $FU_\gamma=U_{\tau(\gamma)}F$.

Proof. Calculating from the definitions:

$$\begin{aligned}F(U_\gamma f)(x)&=\int_{T^2}\exp(-2\pi i\langle x, y\rangle)U_\gamma f(y)\,d\mu(y)\\&=\int_{T^2}\exp(-2\pi i\langle x, \gamma z\rangle)U_\gamma f(\gamma z)\,d\mu(z)\\&=\int_{T^2}\exp(-2\pi i\langle \gamma^* x, z\rangle)f(\gamma^{-1}\gamma z)\,d\mu(z)\\&=Ff(\gamma^*x)=U_{\gamma^{*-1}}Ff(x)=U_{\tau(\gamma)}Ff(x).\end{aligned}$$ Q.E.D.

Note that $\tau(a)=b^{-1}$ and $\tau(b)=a^{-1}$, so many of the expanders in our examples for $SL(2, \mathbf{Z})$ are taken into themselves by τ.

The above proposition implies that $\|M_{T^2,0}\|$ is the same as the norm of the corresponding diffusion operator on the lattice minus the origin: $\mathbf{Z}\times\mathbf{Z}-(0,0)$. The lattice breaks up into disjoint orbits, each characterized by the g.c.d. of the x and y coordinates of its constituents. The isotropy subgroup G_p, where $p=(1,0)$, consists of those matrices having first column $(1,0)$:

$$G_p=\left\{\begin{pmatrix}1 & n\\ 0 & 1\end{pmatrix}\right\}.$$

Since G_p is an Abelian group, it is amenable. We conclude from Proposition 4.1 that the norm of the diffusion on the orbit G-set is the same as the norm of the diffusion on the group G. Hence, we have

PROPOSITION 7.2. $\|M_{T^2,0}\|=\|M\|$.

Gabber and Galil use this bound for the continuous torus to prove expansion results for each discrete torus $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$. The action on $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$ is that induced by the natural action on $\mathbf{Z}\times\mathbf{Z}$. This is the same as that action induced by imbedding $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$ as the set of points in T^2 having coordinates which are both multiples of $1/n$. Divide T^2 up into n^2 little squares, associating $q=(r, s)$ in $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$ with the square

$$\mathrm{Sq}\,(q)=\mathrm{Sq}\,((r, s))=\{(x, y)\in T^2\,|\,r\leqq nx<r+1, s\leqq ny<s+1\}.$$

If γ is an element of G, then the image set $\gamma\,\mathrm{Sq}\,((r, s))$ intersects nontrivially only a few other squares. For example, if $\gamma=A^2=\left(\begin{smallmatrix}1 & 2\\ 0 & 1\end{smallmatrix}\right)$, we have

$$A^2\,\mathrm{Sq}\,((r, s))\subseteq\mathrm{Sq}\,((r+2s, s))\cup\mathrm{Sq}\,((r+2s+1, s))\cup\mathrm{Sq}\,((r+2s+2, s)).$$

Define the degree 12 expander Γ having the following transitions on $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$:

$$\begin{aligned}(x, y)&\to(x\pm(2y+a), y),\\(x, y)&\to(x, y\pm(2x+a))\end{aligned}$$

where $a=0, 1, 2$. Then we have

PROPOSITION 7.3. *The above graph is an $(n^2, 12, \frac{1}{3})$-expander.*

Proof. If X is a subset of $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$, we must show that $|\Gamma(X)|\geqq(1+(1/3n^2)(n^2-|X|))|X|$. Let Sq (Q) denote the subset of T^2 made from little squares corresponding to the points of Q; that is,

$$\mathrm{Sq}\,(Q)=\bigcup_{q\in Q}\mathrm{Sq}\,(q).$$

Denote the four automorphisms A^2, B^2, A^{-2}, B^{-2} of T^2 by $\sigma_1,\cdots,\sigma_4$. Define the spreading operation Γ for subsets of T^2 by $\Gamma(X)=\bigcup_{i=1}^4\sigma_i(X)$. Then we have:

$$\Gamma(\mathrm{Sq}\,(X))\subseteq\mathrm{Sq}\,(\Gamma(X)).$$

But we know that Γ expands subsets of T^2 with expansion rate $\frac{1}{3}$, by a continuous version of Proposition 2.3. Therefore,

$$|\Gamma(X)|=n^2\mu(\mathrm{Sq}\,(\Gamma(X)))\geqq n^2\mu(\Gamma(\mathrm{Sq}\,(X)))$$
$$\geqq n^2\left(1+\frac{1}{3}\mu(\mathrm{Sq}\,(X)^c)\right)\mu(\mathrm{Sq}\,(X))=\left(1+\frac{1}{3n^2}(n^2-|X|)\right)|X|.\qquad\text{Q.E.D.}$$

More generally, if S is a set of elements of $SL(2,\mathbf{Z})$ closed under inversion, it defines an expansion operator on measurable subsets of T^2 by $\Gamma_S(X)=\bigcup_{\sigma\in S}\sigma(X)$. Moreover, for each element σ there will be a finite subset $L_\sigma\subseteq\mathbf{Z}\times\mathbf{Z}$, which can be considered as a set of translations of $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$ for all n, so that independent of n we will have

$$\sigma(\mathrm{Sq}\,(q))\subseteq\mathrm{Sq}\,(\sigma(q)+L_\sigma)=\bigcup_{t\in L_\sigma}\mathrm{Sq}\,(\sigma(q)+t)$$

where addition is performed in the additive group $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$. Actually we should include in L_σ only those t such that the set Sq $(t)\cap\sigma(\mathrm{Sq}\,(0))$ has nonzero measure for all n.

PROPOSITION 7.4. *For each n define the expander Γ_n on $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$ by connecting each point p to all points in the subset $\bigcup_{\sigma\in S}\sigma(p)+L_\sigma$. Then the expansion coefficient for Γ_n is at least as great as that of the spreading operator Γ_S acting on measurable subsets of T^2.*

Another route to the expansion property is to first show how to construct a diffusion operator on $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$ having diffusiveness at least as good as that on T^2, then apply one of the Propositions from § 2. (We did not follow this route in Proposition 7.3 above, because Proposition 2.3 needs to make use of the small value of k.) If the symmetric diffusion M on T^2 is defined by a subset S of G and a probability distribution ϕ on S, with the restriction that $\phi(\sigma)=\phi(\sigma^{-1})$, define the diffusion M_n on $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$ for all n by the random walk that first applies σ with probability $\phi(\sigma)$ then translates by amount $t\in\mathbf{Z}^2$ with probability equal to $\mu((t+\mathrm{Sq})\cap\sigma(\mathrm{Sq}))$ where $\mathrm{Sq}=\{(x,y)\,|\,0\leqq x<1, 0\leqq y<1\}$ is the unit square in $\mathbf{R}^2$.

PROPOSITION 7.5. *For all n we have $\|M_{n,0}\|\leqq\|M\|$.*

Proof. Normalize measures on T^2 and $(\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z})$ to have total mass 1. Define the orthonormal projection map P from $L^2(T^2)$ to $L^2((\mathbf{Z}/n\mathbf{Z})\times(\mathbf{Z}/n\mathbf{Z}))$ by

$$P(f)(q)\equiv n^2\int_{x\in Sq(q)}f(x)\,d\mu(x).$$

Define the isometric injection I the other way by $I(f)(x)\equiv f(q)$ when $x\in\mathrm{Sq}(q)$. Then $M_{n,0}=PM_{T^2,0}I$ implies

$$\|M_{n,0}\|\leqq\|P\|\cdot\|M_{T^2,0}\|\cdot\|I\|=\|M\|.\qquad\text{Q.E.D.}$$

8. Bounded concentrators built from expanders and superconcentrator networks built from bounded concentrators. See Gabber-Galil [11].

DEFINITION. For $\theta < 1$, an (n, θ, k, α) *bounded concentrator* is a bipartite graph Γ with n inputs, θn outputs, at most kn edges, such that every input subset X with $|X| \leqq \alpha n$ maps to an output set $\Gamma(X)$ at least as large.

Philip Hall's Marriage Lemma implies that any (n, θ, k, α) bounded concentrator is actually a *concentrator* in the following sense: for any input subset X with $|X| \leqq \alpha n$ there will be a set of $|X|$ disjoint edges connecting X to an equal number of outputs.

Gabber and Galil use an (m, k, d) expander to make an $(n = m(p+1)/p,\ \theta = p/(p+1),\ k',\ \alpha = 1/2)$ bounded concentrator, where $k' = (k+1)p/(p+1)$. Their construction proceeds as follows.

Construction. Assume that p divides m evenly. Break the input set into a big part of size m and a little part of size m/p. Use an (m, k, d) expander to connect the big part of the input set to the output set (which has size m). Divide the output set into m/p clumps of size p, and connect each of the inputs in the small set to all of the members of one of the output clumps, each input to a different clump. The resulting bipartite graph works, as long as $d \geqq 2p^2/(p-1)(p^2+1)$.

Proof. Let k be the number of inputs in a given subset, with $k \leqq m(p+1)/2p$. We need to show that these connect to a total of at least k outputs. If at least k/p of these appear in the small set, then we are done, because $p \cdot k/p \geqq k$. Otherwise, there will be at least $q = k(p-1)/p$ of these in the large input set. Since these feed the expander, we will succeed if the expansion factor is at least $p/(p-1)$; namely, if $1 + d(1 - q/m) \geqq p/(p-1)$. But

$$1 + d\left(1 - \frac{q}{m}\right) = 1 + d\left(1 - \frac{k(p-1)}{pm}\right) \geqq 1 + d\left(1 - \frac{p^2-1}{2p^2}\right)$$

and the last is large enough when $d \geqq 2p^2/(p-1)(p^2+1)$. Q.E.D.

Of course, the condition $d \geqq 2/(p-1)$ will also suffice.

DEFINITION. An (n, k) *superconcentrator* is a directed acyclic graph with n inputs and n outputs, having at most kn edges, such that for any choice of r inputs and r outputs there is a collection of r disjoint paths starting in the input set and ending in the output set.

Pippenger [18] gives a recursive construction of superconcentrators in terms of bounded concentrators, which is generalized by Gabber and Galil as follows. Connect input i directly to output i for $1 \leqq i \leqq n$, using n edges. Given an input subset and an equinumerous output subset, use these direct lines, if possible, to connect inputs to outputs by paths of length 1. After that there remain fewer than $n/2$ unmatched inputs to be connected to a set of unmatched outputs. Use an $(n, \theta, k, \frac{1}{2})$ bounded concentrator to concentrate the unmatched inputs into a new "input space" of size θn. Similarly concentrate the unmatched outputs into an "output space" of size θn. Then use a size θn superconcentrator to match the unmatched. Start the recursive definition by using, say, a complete n^2 edge graph as superconcentrator when n is small. The result of this construction is summarized in Gabber-Galil's Lemma 8:

LEMMA (from [11]). *If we can construct for all n and $(n, \theta_n, k, 1/2)$ bounded concentrator, where $\theta_n = \theta + \varepsilon_n$, $0 \leqq \theta < 1$, $\varepsilon_n = o(1)$, and $k > 1$, then we can construct a family of linear superconcentrators of density $(2k+1)/(1-\theta)$.*

A consequence of this is Gabber-Galil's Theorem 3:

THEOREM (from [11]). *Assuming we can construct for every $n = m^2$ an $(n, k, 2/(p-1))$ expander, then we can construct a family of linear superconcentrators of density $(2k+3)p+1$.*

For example, we have already demonstrated a family of expanders with $n=m^2$, $k=12$, and $d=\frac{1}{3}$. Since $\frac{1}{3}=2/(7-1)$ we can choose $p=7$ in the above Theorem, obtaining a family of superconcentrators of density $(2\cdot 12+3)\cdot 7+1=190$. By using the inequality appearing here as Proposition 2.5, Alon and Milman in [5], [6], and with Galil in [4], turned the degree 12 diffusers into degree 13 expanders and made superconcentrators of density 157.35. We provide evidence in § 9, Ex. 2 for the existence of $k=4$ diffusers having $d=\frac{1}{3}$. If these were sufficiently numerous, we would be able to build superconcentrators of density $(2\cdot 4+3)\cdot 7+1=78$.

Gabber-Galil's $k=7$ example has $d\geqq\frac{1}{6}$ so we can choose $p=13$ and build superconcentrators with density $(2\cdot 7+3)\cdot 13+1=222$.

Alon suggested the use of the Alon-Milman isoperimetric inequality (Proposition 2.5), instead of Proposition 2.2, for making expanders with $k\geqq 5$. Conjectural families of optimal degree $k-1$ diffusers would have $\|M_0\|\leqq 2\sqrt{k-2}/(k-1)$, and would be used to make expanders of degree k. Looking back at the proof of the bounded concentrator construction, we see that the active input subset in the "big" part must expand by the ratio at least $p/(p-1)$. Define p_3 to be the smallest real number so that the expansion predicted by Proposition 2.5 gives this ratio. If Gabber-Galil's construction were to work for nonintegral values of p, we would achieve superconcentrators of density $k_3=(2k+3)p_3+1$. The construction described in their Appendix 1 allows the construction of bounded concentrators with effective $p=p_2$ and superconcentrator density k_2. We denote by p_1 the best integer value and by k_1 the density of the corresponding superconcentrator. In Table 1 we display the results for $5\leqq k\leqq 13$. (Outside of this range Proposition 2.2 leads to slightly better results.) Notice that k_2 are the densities that we know how to achieve, given the existence of the required expanders. For $k=8$ we would have superconcentrators of density 57.1587.

TABLE 1

k	p_1	k_1	p_2	k_2	p_3	k_3
5	6	79	5.3915	71.0890	5.0945	67.2280
6	4	61	3.8762	59.1430	3.8208	58.3125
7	4	69	3.4471	59.6004	3.2568	56.3657
8	3	58	2.9557	57.1587	2.9397	56.8534
9	3	64	2.7998	59.7963	2.7361	58.4585
10	3	70	2.6847	62.7491	2.5941	60.6636
11	3	76	2.5962	65.9041	2.4890	63.2258
12	3	82	2.5258	69.1953	2.4080	66.0164
13	3	88	2.4683	72.5817	2.3435	68.9609

9. Numerical results for diffusers.

Example 1. ($k=3$) For a prime p take the vertex set to be the projective line $P^1(\mathbf{F}_p)$. An inhomogeneous coordinate z takes on values $0, 1, 2, \cdots, p-1, \infty$. A matrix in $SL(2, \mathbf{Z})$ acts as a linear fractional transformation sending z to $(az+b)/(cz+d)$.

From the action of the elements C, DCD, $D^{-1}CD^{-1}$ we obtain the three permutations (expressed as functions of z):

$$-z^{-1}, (2-z)^{-1}, 2-z^{-1}$$

(where, of course, $1/0=\infty$ and $1/\infty=0$). Giving each of these three transitions equal probability, we define a Markov process M on $P^1(\mathbf{F}_p)$. Computer experiments show that $\|M_0\| \leqq 2\sqrt{2}/3 \approx 0.9428090$ for all primes less than 700 with the exception of $p=433$, 479.

Example 2. ($k=4$) From the action of the elements A^2, A^{-2}, B^2, B^{-2} on the projective line $P^1(\mathbf{F}_p)$ we obtain the four permutations

$$z+2, z-2, \frac{1}{z^{-1}+2}, \frac{1}{z^{-1}-2}.$$

Give each of these four transitions equal probability. Computer experiments show that $\|M_0\| \leqq \sqrt{3}/2 \approx 0.8660254$ for all primes less than 700 with the exception of $p=251$, 331, 461, 479, 569, 617.

Example 3. ($k=3$) From the action of the elements C, D, D^{-1} on the projective line we obtain the three permutations

$$-z^{-1}, (1-z)^{-1}, 1-z^{-1}.$$

The norm of the master diffusion acting on the infinite group $SL(2, \mathbf{Z})$ is $(1+\sqrt{8\sqrt{2}+13})/6 \approx 0.988482$ by the work in § 5. For primes less than 700 only $p=433$, 479 fail to have $\|M_0\|$ less than this bound.

Example 4. ($k=4$) From the action of the elements D, D^{-1}, CDC, $CD^{-1}C$ on the line we obtain the four permutations

$$(1-z)^{-1}, 1-z^{-1}, -1-z^{-1}, -(1+z)^{-1}.$$

Again numerical experiments show that for all of the 124 odd primes less than 700 except for $p=433$, 479 we have $\|M_0\| \leqq (2\sqrt{2}+1)/4 \approx 0.957107$. These are notably the same exceptions as for Examples 1 and 3. This example is equivalent to the one using the transitions given by the four elements A, B, A^{-1}, B^{-1}.

Example 5. ($k=6$) The three elements A^4, A^2B^2, B^4 generate a free subgroup of $SL(2, \mathbf{Z})$. The norm of the diffusion there is precisely $\sqrt{5}/3$. For the diffusion acting on the finite sets $P^1(\mathbf{F}_p)$ computer experiments show that for the 124 odd primes less than 700 we have $\|M_0\| \leqq \sqrt{5}/3 \approx 0.745356$ except for the following 18 primes: 41, 61, 103, 107, 173, 179, 251, 337, 379, 421, 461, 479, 577, 593, 617, 641, 661, 677.

Example 6. ($k=8$) Whenever two elements α, β freely generate a subgroup, the four elements $\alpha\beta$, $\alpha^2\beta^2$, $\alpha^3\beta^3$, $\alpha^4\beta^4$ will freely generate a subgroup. Let $\alpha=A^2$ and $\beta=B^2$. The norm of the diffusion operator on the infinite group will be $\sqrt{7}/4$. Acting on the finite projective lines, however, $\|M_0\|$ exceeds $\sqrt{7}/4$ often: for 49 of the 124 odd primes less than 700, although the values may still converge to $\sqrt{7}/4 \approx 0.6614378$ as the primes go to ∞. For $p=677$ we obtained 0.668479.

Example 7. ($k=4$) Since random constructions are known to work well for bounded concentrators and in fact work better (so far) than explicit constructions, we suggest the following random method for making diffusers. Let $\mathbf{Z}/n\mathbf{Z}$ be the vertex set. Choose two n-cycles σ_1, σ_2 at random. Allow the four transitions

$$\sigma_1^{-1}(z), \sigma_1(z), \sigma_2^{-1}(z), \sigma_2(z)$$

with equal probability. Of course one can choose σ_1 to be $z \to z+1$. For each of the 124 odd primes p less than 700 we generated a random diffuser on the points of $\mathbf{Z}/p\mathbf{Z}$ and found that the second largest eigenvalue modulus exceeded $\sqrt{3}/2$ only 24 times.

REFERENCES

[1] M. Ajtai, J. Komlós and E. Szemerédi, *Sorting in c* log *n parallel steps*, Combinatorica, 3 (1983), pp. 1-19.

[2] ———, *An $O(n \log n)$ sorting network*, Proc. 15th Annual ACM Symposium on Theory of Computing, Boston, 1983, pp. 1-9.

[3] N. Alon, *Eigenvalues and expanders*, preprint, 1984.

[4] N. Alon, Z. Galil and V. D. Milman, *Better expanders and superconcentrators*, preprint, 1984.

[5] N. Alon and V. D. Milman, *Eigenvalues, expanders and superconcentrators*, Proc. 25th Annual Symposium on Foundations of Computer Science, Gainesville, Florida, 1984, pp. 320-322.

[6] ———, λ_1, *isoperimetric inequalities for graphs, and superconcentrators*, J. Combin. Theory Ser. B, 38 (1985), pp. 73-88.

[7] L. A. Bassalygo, *Asymptotically optimal switching circuits*, Problemy Peredachi Informatsii, 17(3), (1981), pp. 81-88; Problems Inform. Transmission, 17(3) (1981), pp. 206-211.

[8] L. A. Bassalygo and M. S. Pinsker, *Complexity of an optimum non-blocking switching network without reconnections*, Problemy Peredachi Informatsii, 9(1) (1973), pp. 84-87; Problems Inform. Transmission, 9(1) (1973), pp. 64-66.

[9] F. R. K. Chung, *On concentrators, superconcentrators, generalizers, and nonblocking networks*, Bell System Tech. J., 58 (1978), pp. 1765-1777.

[10] M. M. Day, *Convolutions, means, and spectra*, Illinois J. Math., 8 (1964), pp. 100-111.

[11] O. Gabber and Z. Galil, *Explicit constructions of linear-sized superconcentrators*, J. Comput. System Sci., 22 (1981), pp. 407-420.

[12] E. Hewitt and K. Ross, *Abstract Harmonic Analysis*, Vol. 1, Springer-Verlag, Berlin, Göttingen, Heidelberg, 1963.

[13] H. Kesten, *Symmetric random walks on groups*, Trans. Amer. Math. Soc., 92 (1959), pp. 336-354.

[14] ———, *Full Banach mean values on countable groups*, Math. Scand., 7 (1959), pp. 146-156.

[15] M. Klawe, *Limitations on explicit constructions of expanding graphs*, SIAM J. Comput., 13 (1984), pp. 156-166.

[16] G. A. Margulis, *Explicit constructions of concentrators*, Problemy Peredachi Informatsii, 9(4) (1973), pp. 71-80; Problems Inform. Transmission, 9(4) (1973), pp. 325-332.

[17] M. S. Pinsker, *On the complexity of a concentrator*, Proc. 7th International Teletraffic Conference, Stockholm, June 1973, pp. 318/1-318/4.

[18] N. Pippenger, *Superconcentrators*, SIAM J. Comput., 6 (1977), pp. 298-304.

[19] R. M. Tanner, *Explicit concentrators from generalized N-gons*, this Journal, 5 (1984), pp. 287-293.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

014

CHARACTERIZATION AND RECOGNITION OF PARTIAL 3-TREES*

STEFAN ARNBORG† AND ANDRZEJ PROSKUROWSKI†‡

Abstract. Our interest in the class of k-trees and their partial graphs and subgraphs is motivated by some practical questions about the reliability of communication networks in the presence of constrained line- and site-failures, and about the complexity of queries in a data base system. We have found a set of confluent graph reductions such that any graph can be reduced to the empty graph if and only if it is a subgraph of a 3-tree. This set of reductions yields a polynomial time algorithm for deciding if a given graph is a partial 3-tree and for finding one of its embeddings in a 3-tree when such an embedding exists. Our result generalizes a previously known recognition algorithm for partial 2-trees (series-parallel graphs).

Key words. graph reductions, confluent reductions, k-trees

AMS(MOS) subject classification. 05C10

1. Introduction. Our interest in the class of k-trees and their subgraphs is motivated by some practical questions about the reliability of communication networks in the presence of constrained line—and site—failures (Farley [5], Farley and Proskurowski [7], Neufeld and Colbourn [10], Wald and Colbourn [14]) and about the complexity of queries in a data base system (Arnborg [1]).

We will briefly describe the connection between the problem of finding the minimal value of k for which a given graph is a partial k-tree and the complexity of queries. Let us consider conjunctive data base queries, an important class of queries from which answers to less restrictive classes of queries can be constructed. Such a query has the form

$$\bigwedge_{1\leq i\leq k} P_i(x_{i,1}, x_{i,2}, \cdots, x_{i,n_i}).$$

The variables occurring in the relation P_i constitute a subset of n variables $x_1, x_2, \cdots, x_n$; the relation is a set of n_i-tuples of values from a common domain (or from several domains). The query asks if there is an assignment of values $a_1, \cdots, a_n$ to the n variables so that the tuple $\langle a_{i,1}, a_{i,2}, \cdots, a_{i,n_i}\rangle$ belongs to P_i, for each i. The cost of a conjunctive query involving only two relations depends critically on the size of the relations (the number of tuples satisfying them). For a more complex query, minimization of sizes of intermediate relations by means of variable elimination is of great import. This is achieved by answering a partial conjunctive query involving all relations containing a given variable. After this join, the variable can be eliminated from further consideration by simple projection. Joining the relations until only one relation remains evaluates the conjunctive query. If this final relation is nonempty, then the answer is "yes." The size of a k-ary relation may be as large as m^k, where m is the size of the domain. Thus, a relevant objective function for finding the best join order is the maximum arity (the number of variables) of an intermediate relation. Minimization of this objective function is equivalent to embedding a graph obtained from the query syntax into a $(k-1)$-tree, for the minimum k.

In the remainder of § 1 we introduce some standard graph terminology and reduction operations on graphs. In § 2 we review some properties of k-trees and

* Received by the editors January 24, 1984, and in revised form November 15, 1984.

† Department of Numerical Analysis and Computing Science, The Royal Institute of Technology, S-100 44 Stockholm 70, Sweden.

‡ On leave from Computer and Information Science Department, University of Oregon, Eugene, Oregon 97403.

introduce the class of k-decomposable graphs, which is then shown identical to the class of partial graphs of k-trees. In § 3 we exhibit a set of reduction operations such that a graph is reduced to the empty graph by the rules *iff* it is a partial 3-tree. The rules are confluent (or, equivalently, have the Church–Rosser property), which means that the reductions can be applied in any order.

We will consider simple, loopless, undirected combinatorial graphs. Two vertices u and v of a graph G are called *adjacent* if there is an edge (u, v) of G; the edge (u, v) is said to be *incident* with its *end-vertices* u and v. The set of all vertices adjacent to a given vertex v is called the (*open*) *neighborhood* of v in $G, \Gamma_G(v)$ or $\Gamma(v)$ when G is clear from context. The order of $\Gamma(v)$ is called v's *degree*, and its elements are called v's *neighbors*. For a given graph G with the vertex set V and the edge set E, we define a *subgraph induced* by a subset U of vertices to be a graph with the vertex set U and the edge set D of all edges of G with both end-vertices in U. A *clique* is a completely connected subgraph. A *partial graph* of G is defined as a graph with the vertex set V and edge set D, a subset of E. A *subgraph* of G is a partial graph of an induced subgraph of G.

We will investigate classes of graphs which can be defined by the following operators on graphs (see also Rose, Tarjan and Lueker [13]):

Star substitution, $S_k(G, v) = H$, where v is a vertex of G of degree k; the vertex set of H is $V - \{v\}$ and its edge set is $E - \{(u, v) | u \in \Gamma(v)\} \cup \{(u, w) | u, w \in \Gamma(v)\}$. (A star centered in v is "substituted" by a complete graph defined by its neighbors.) This is the vertex elimination operation of Rose [11].

Isolated vertex removal, $I(G, v) = H$, where v is an isolated vertex (with no incident edges); H has the same edge set as G and its vertex set is $V - \{v\}$.

Star removal, $R_k(G, v) = H$, where v is a vertex of G and the subgraph of G induced by $\Gamma(v)$ is a complete graph with k vertices; then $H = S_k(G, v)$.

Star hook-up $H_k(G, K) = H$, where K is a clique induced by k vertices of G; H has the vertex set $V \cup \{w\}$, $w \notin V$, and the edge set $E \cup \{(u, w) | u$ is a vertex of $K\}$.

Extended operators, S'_k and R'_k, are defined as the unions of I and S_i and R_i, respectively, for all i between 1 and k.

We define the corresponding relations $\mathbf{S}_k$, $\mathbf{R}_k$, $\mathbf{H}_k$, $\mathbf{S}'_k$, and $\mathbf{R}'_k$ to hold between two graphs G and H iff there exists an element of G (a vertex or a clique) so that H is the result of applying the corresponding operator to G and its element. Finally, we define the class of k-trees, $\mathcal{T}_k$, (cf. Beineke and Pippert [4] and Rose [11], [12]) as the family of graphs for which the (reflexive) transitive closure of the relation $\mathbf{H}_k$, $\mathbf{H}_k^*$, holds with K_k, the complete graph with k vertices. A *k-leaf* is a vertex of degree k in a k-tree (and in a 3-tree we similarly have 3-leaves). It follows straightforwardly from the definition that a k-tree has at least two k-leaves, and that k-leaves are nonadjacent in a k-tree with more than $k+1$ vertices. The class of *partial k-trees*, $\mathcal{PT}_k$, is defined to consist of all subgraphs of k-trees (cf. Wald and Colbourn [14], for the case of $k = 2$).

It might be interesting to view the reductions S_k as simplifying rewrite rules (this viewpoint has been taken by Liu and Geldmacher [9] who considered the implications of series-parallel reductions). In that context, we would be looking for a set of reduction rules confluent under a congruence relation for which the class of partial k-trees is an equivalence class. Here, the set of reduction rules is confluent if for any two graphs resulting from reducing a given graph in different ways, there exists a reduct graph reachable by reductions from either of the two graphs. Since every chain of reductions of a finite graph is finite, this global confluence property is implied by local confluence, where we require the existence of a common reduct for graphs differing only by one application of different reduction rules (the so-called "diamond lemma," see for instance Huet and Oppen [8]).

2. General k-trees. The following two theorems are equivalent to the example stated in Rose [11] implying that, for a k-tree G, there is a sequence of star removals leading to a complete graph with k vertices, and that any star removal can start this sequence.

THEOREM 2.1. *G is a k-tree iff $G\mathbf{R}_k^* K_k$.*

THEOREM 2.2. *G is a k-tree iff either G is K_k or every H such that $G\mathbf{R}_k H$ is a k-tree.*

The latter gives the basis for an obvious k-tree recognition algorithm (cf. Rose [11]): given a graph G, iteratively remove vertices of degree k with completely adjacent neighbors until no further removal is possible; G is a k-tree iff the remaining graph is the K_k.

After a couple of technical lemmas, we will present a theorem stating that a graph is a partial k-tree if and only if there is a sequence of reductions S'_k which leads to a reduction to the empty graph (with no vertices).

LEMMA 2.3. *For every complete subgraph S of i vertices in a k-tree G $(i<k)$, there exists a complete subgraph K of k vertices in G of which S is an induced subgraph.*

Proof. G can be reduced by a series of applications of operation R_k so that in the resulting k-tree H, no vertex of S is removed and at least one vertex v of S has degree k. In H, the vertex v has a completely connected neighborhood K, or else there would be no way of reducing H to a complete graph through a series of star removals, R_k. K contains all vertices of S. □

LEMMA 2.4. *Any graph with not more than k vertices is a partial k-tree.*

Proof. The complete graph K_k can be constructed by adding, if necessary, new vertices and missing edges to the original graph. □

THEOREM 2.5. *G is a partial k-tree iff $G\mathbf{S}_k'^*\varnothing$.*

Proof. ($\rightarrow$, by induction on the order of a k-tree H containing G as a subgraph.) The basis follows by Lemma 2.4 since any vertex v of such a graph G can be chosen for a reduction $S'_k(G, v)$ which, repeated, leads to the empty graph. Let us assume that for any graph G which is a subgraph of a k-tree H with n or less vertices, $G\mathbf{S}_k'^*\varnothing$. Consider a graph G which is a subgraph of a k-tree H with $n+1$ vertices. H has a vertex v of degree k with completely connected neighbors (Rose [11]). If v is a vertex of G, then $G' = S'_k(G, v)$ is a subgraph of $H-\{v\}$; otherwise G is a subgraph of $H-\{v\} = R_k(H, v)$. It follows from the inductive assumption that $G\mathbf{S}_k'^*\varnothing$.

($\leftarrow$, by induction on the order of G.) By Lemma 2.4, we need to consider only graphs with at least k vertices. Let us assume that all graphs with n or fewer vertices which can be reduced to the empty graph by a series of S'_k reductions are partial k-trees. Consider a graph G with $n+1$ vertices and such that $G\mathbf{S}_k'^*\varnothing$. Let v be the vertex of G which is removed in the first of these reductions. Thus, by the inductive assumption, $S'_k(G, v)$ is a subgraph of some k-tree H. By Lemma 2.3, the neighborhood of v is contained in a k-complete subgraph of K and H. Applying the hook-up operation to H and K results in a k-tree containing as a subgraph a graph is isomorphic to G, the new vertex w corresponding to v in G. □

Partial k-trees can be embedded in k-trees without adding any new vertices:

THEOREM 2.6. *Any partial k-tree with at least k vertices can be completed to a k-tree with the same number of vertices.*

Proof (by induction on the number of vertices of G). The theorem is obviously true for G with k vertices. Assume that it is true for all partial k-trees with n vertices and consider a partial k-tree G with $n+1$ vertices. Let v be a vertex of G of degree not greater than k such that the graph $G' = S'_k(G, v)$ in a partial k-tree. There is a *k-tree* H with n vertices, with G' as a partial graph, and such that a k-complete subgraph K of H contains all vertices of $\Gamma(v)$ (see Lemma 2.3). The k-tree $H_k(H, K)$ has $n+1$ vertices and contains G as its partial graph. □

An alternative definition of partial k-trees can be given using the notion of k-decomposability: 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.

THEOREM 2.7. *The class of k-decomposable graphs is exactly $\mathcal{PT}_k$.*

Proof. Since all minimal separators in a k-tree have order k (Beineke and Pippert [4]), k-trees are k-decomposable, together with all their partial graphs. If a k-decomposable graph with more than $k+1$ vertices can be embedded in a union of two k-trees with at most k completely connected vertices in common, then—by Lemma 2.3—the common part can be extended to a k-complete graph, the embedding graph can be extended to a k-tree, and the theorem follows by induction on the order of the graph. □

Unfortunately, this characterization of partial k-trees does not give us an efficient algorithm for recognition of this class of graphs since the separator property is lost in subgraphs. Namely, in a partial k-tree we may be able to find a "small" separator which cannot be extended to a complete subgraph of a k-tree in an embedding of the partial k-tree.

3. Partial 3-trees. Wald and Colbourn [14] restate Duffin's [5] characterization of series-parallel graphs by completely characterizing the class of partial 2-trees as graphs with no subgraphs homeomorphic to K_4. This characterization does not carry into higher values of k. Figure 1(a) shows a planar graph (which cannot have a homeomorph of K_5) which is not a partial 3-tree.

A natural generalization of the recognition algorithm for series-parallel graphs [5], [10] would be to perform applicable reductions in S_3' in any sequence. Since the operations I and S_1 do not introduce any new edges, they result in partial 3-trees whenever applied to a partial 3-tree. However, the other reduction operations in S_3' may not be "safe," i.e., a partial 3-tree may be reduced to a graph which is not a partial 3-tree. An example is given in Fig. 1(b), where a partial 3-tree can be reduced to the graph in Fig. 1(a) by application of S_3 to vertex v. We can recognize a simple case of safe application of the reduction S_3':

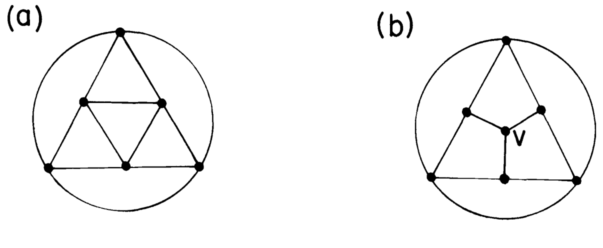


FIG. 1. (a) *The* 6*-vertex,* 4*-regular plane graph G, and* (b) *a partial* 3*-tree H such that $G=S_3(H, v)$.*

THEOREM 3.1. *For any partial* 3*-tree G, $S_2(G, w)$ is a partial* 3*-tree and $S_3(G, v)$ is a partial* 3*-tree if at least two neighbors of v are adjacent. See Fig.* 2.

Proof (by induction on the number of vertices in G). The theorem is obviously true if G has not more than 3 vertices. Assume that in any partial 3-tree with less than $n>3$ vertices such reductions lead to partial 3-trees. Consider a partial 3-tree G with n vertices and a reduction $G_3'(G, u)$ resulting in a partial 3-tree, G'. If $u=w$, $u=v$, or u is not any of the neighbors of v or w, then the theorem follows directly from the assumption. If degree of u is less than 3, the cases of its adjacencies are trivial. Otherwise, there are three cases to consider: (i) u is a neighbor of w. Then, $S_3(G', w)$

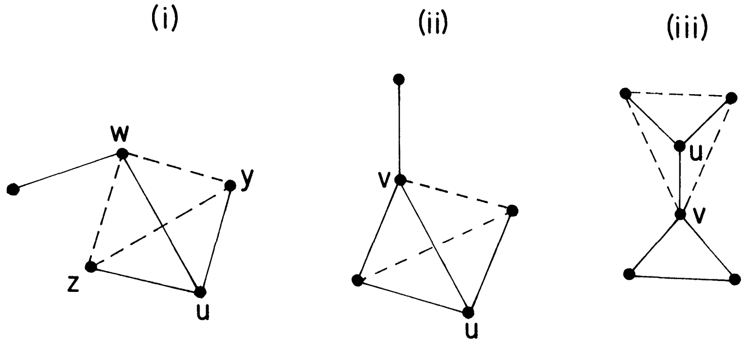


FIG. 2. *Cases in the proof of Theorem* 3.1.

is a partial 3-tree by the inductive assumption (two of w's neighbors in G', y and z, are adjacent) and it is isomorphic to $S_3(S_2(G, w), u)$. (ii) u is one of the adjacent neighbors of v. If u is also adjacent to the third neighbor of v, then G' is isomorphic to $S_3(G, v)$. Otherwise, $S_3(G', v)$ is a partial 3-tree by the inductive assumption (v's neighbors in G', x and y, are adjacent) and it is isomorphic to $S_3(S_3(G, v), u)$. (iii) u is not adjacent to the other two neighbors of v. Then, $S_3(G, v)$ is isomorphic to a partial graph of $S_3(G, u)$, which is a partial 3-tree by the assumption. This completes the proof of the inductive step. □

The operation S_3 plays a crucial role in the eventual reduction of vertex degrees in the graph, even though the star substitution operation applied to vertices with independent (nonadjacent) neighbors increases their degree. Fortunately, we are able to isolate configurations involving such independent neighborhoods of vertices of degree 3 that make the degree reduction possible.

THEOREM 3.2. *A graph G without vertices of degree* 0, 1, *or* 2, *and with no vertex of degree* 3 *that has two adjacent neighbors is a partial* 3*-tree only if there are subgraphs of G isomorphic to either C' or C'' in Fig.* 3, *where vertices u, v and w have degree* 3 *in G, and vertex x of C' has degree* 3 *in G.*

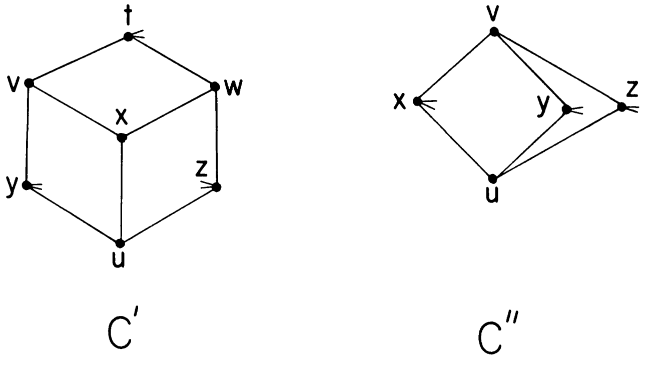


FIG. 3. *The necessary subgraphs in a partial* 3*-tree.*

Proof. Consider a partial 3-tree G such that its minimum vertex degree is 3 and no two neighbors of a degree 3 vertex are adjacent. Let H be one of G's embedding 3-trees. H' is the induced subgraph of H obtained by deleting all 3-leaves from H (at least two exist). G' is the graph obtained from G by removing the 3-leaves with the S_3 reduction. G' is a partial graph of H'. Let x be any 3-leaf of H' (or any vertex if

$H'=K_4$), and let L_x be the nonempty set of 3-leaves of H adjacent to x. Every member of L_x has degree 3 in H and thus degree 3 or less in G which is a partial graph of H. But G has also minimum degree 3, so each member of L_x has degree 3 in G. Since no two neighbors of a degree 3 vertex are adjacent in G, no two neighbors of a vertex in L_x are adjacent in G, so L_x and x have disjoint neighborhoods in G. The degree of x in G' is not greater than 3 and x's neighborhood in G' consists of two disjoint sets, $\Gamma_G(x)-L_x$ (the original neighbors) and $\Gamma_G(L_x)-\{x\}$ (the neighbors introduced by S_3 transformations), i.e.:

$$|\Gamma_G(x)-L_x|+|\Gamma_G(L_x)-\{x\}|=|\Gamma_{G'}(x)|\leqq 3.$$

The possible solutions to this inequality are severely constrained by the degree assumption on G and the fact that for both set differences the second operand is a subset of the first. Since G has minimum vertex degree 3, the second term in the left-hand side is at least 2, so the first term can only be 0 or 1, and $|L_x|$ is at least two. We split cases by the second term, which can be 2 or 3:

(i) $|\Gamma_G(L_x)|=3$. All vertices in L_x have the same neighbors, and since they are at least two, configuration C'' must be present.

(ii) $|\Gamma_G(L_x)|=4$. This case implies $\Gamma_G(x)=L_x$, so $|L_x|$ must be at least 3, and every vertex in L_x is G-adjacent to x and two other vertices among the three in $\Gamma_G(L_x)-\{x\}$. This implies configuration C'' if $|L_x|\geqq 4$ and C' or C'' if $|L_x|=3$. Since $\Gamma_G(x)=L_x$, whenever only configuration C' is present, one of its occurrences must have degree 3 in G for its vertex x. □

Now that we know of the necessity of subgraphs with vertices of degree 3 involved in triangles or squares, we have to establish safe reductions thereof. For example, in the graph G (see Fig. 4) which has the graph C' of Fig. 3 as a subgraph, reduction $S_3(G,x)$ leads to a graph which is not a partial 3-tree even though the original graph G is one.

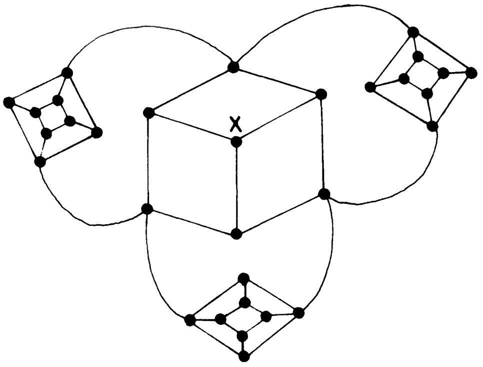


FIG. 4. *A partial 3-tree G with an unsafe reduction $S_3(G,x)$.*

THEOREM 3.3. *For any partial 3-tree G with a subgraph isomorphic to either graph C' or C'' in Fig. 3, the graphs $S_3(G,u)$, $S_3(G,v)$, and $S_3(G,w)$ are all partial 3-trees if vertices u, v, and w have all degree 3 in G.*

Proof (by induction on the number of vertices of G). By inspection, the thesis is true for graphs with no more than 6 vertices. Assume it is true for graphs with fewer than $n>6$ vertices and consider a partial 3-tree G with n vertices. By Theorem 2.5, there is a vertex s such that $S_3'(G,s)$ is a partial 3-tree. If s is one of the vertices u, v or w, or is not adjacent to any of them, then the thesis follows by the inductive assumption. Otherwise, we have to consider three cases: (i) In a subgraph isomorphic to C', $s=x$. Applying the operation S_3 to u, v, w and x, in this order, reduces graph G to a graph G_1, where the three remaining vertices induce a triangle (see Fig. 5(a)). Since the graph $S_3(G, s)$ is a partial 3-tree, so is its subgraph G_2 in which the edges (u, v), (v, w), and (u, w) are missing. But G_2 is reducible to G_1 by application of the operation S_2 (which is safe by Theorem 3.1) to u, v, and w (see Fig. 5(b)). Thus, $S_3(G, u)$ is a partial 3-tree.

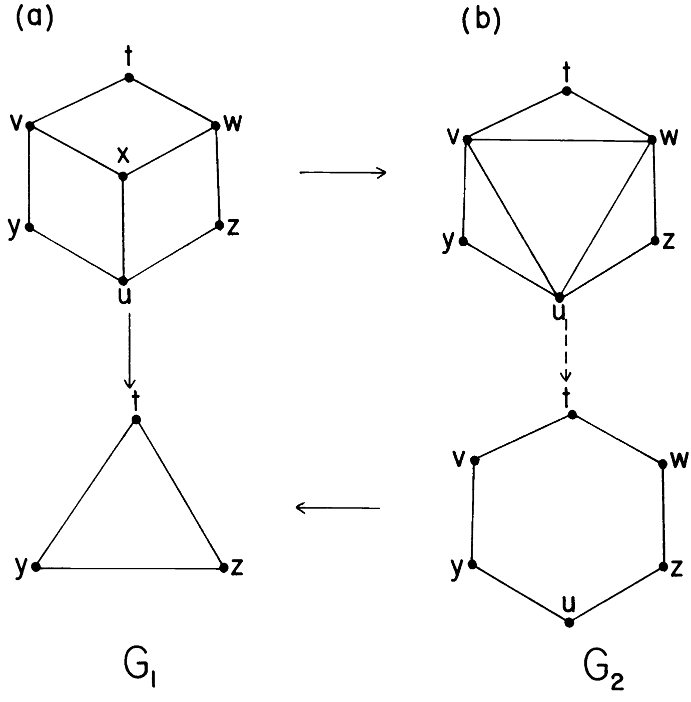


FIG. 5. *Reduction of a partial* 3*-tree* G, *case* (i).

(ii) In a subgraph isomorphic to C', $s=y$ (by symmetry, a similar argument holds for $s=z$ and $s=t$). Applying the operation S_3 to vertices u, v, and w in this order reduces G to a graph containing as a subgraph the graph G_1 (see Fig. 6(a)). On the other hand, since $S_3(G, y)$ is a partial 3-tree, so is its subgraph G_2 in which the edge (q, v) is missing (q is the third neighbor of y). But G_2 is reducible to a graph isomorphic to G_1 (in which u is isomorphic to y in G_1) by a sequence of safe applications of the operation S_3 to v and w (see Fig. 6(b) and Theorem 3.1). Thus, $S_3(G, u)$ is a partial 3-tree.

(iii) In a subgraph isomorphic to C'', $s=x$ (by symmetry, a similar argument applies to the cases $s=y$ or $s=z$). Applying the operation S_3 to u, v, and x in this order reduces G to a graph G_1 where the remaining vertices induce a triangle (see Fig. 7(a)). The graph $G_2=S_3(G, s)-\{(v, q)\}$ (where q is the third neighbor of x) is a partial 3-tree and is reducible to G_1 by application of the safe instances of the operation S_3 to v and u in this order (see Fig. 7(b)). Thus G_1 is a partial 3-tree. □

It should be obvious that the reduction S_2' is confluent in a system recognizing partial 3-trees, since $S_2'(S_2'(G, u), v)=S_2'(S_2'(G, v), u)$, for any graph G and its two vertices u and v. By inspection of cases when two adjacent vertices can be reduced according to any two safe reduction rules, we can easily show that the set of instances of S_3' investigated in Theorems 3.1 and 3.3 are confluent.

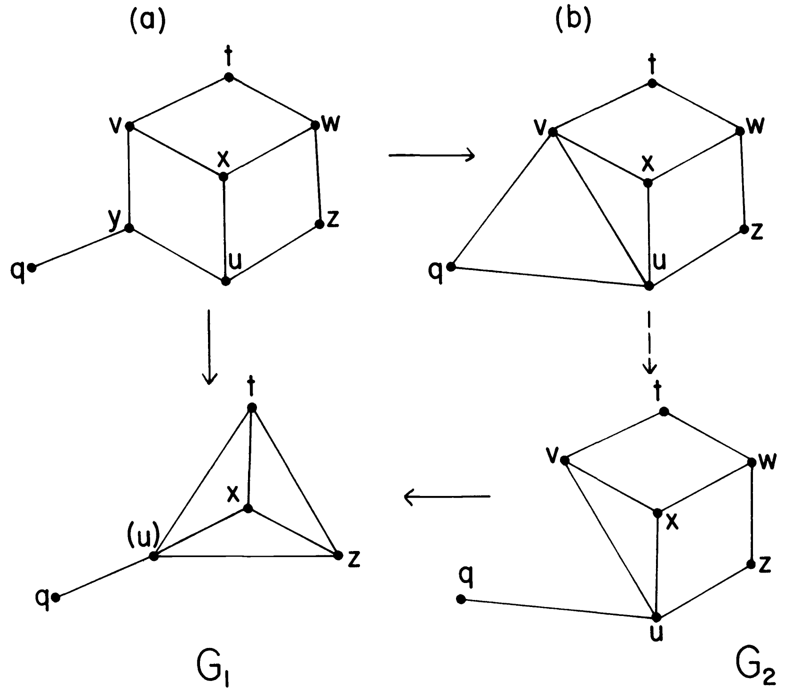


FIG. 6. *Reduction of a partial* 3*-tree G, case* (ii).

THEOREM 3.4. *The following reduction rules are confluent under a congruence relation under which all partial* 3*-trees are equivalent to the empty graph*: *isolated vertex removal, reduction of a vertex of degree* 1, *reduction of a vertex of degree* 2, *and star-triangle substitution when* (i) *two of the neighbors of a vertex of degree* 3 *are adjacent* (*Theorem* 3.1), (ii) *all the neighbors of a vertex of degree* 3 *are also neighbors of one other vertex of degree* 3, *or* (iii) *the neighbors of a vertex of degree* 3 *are shared with those of two other vertices of degree* 3 *that also share a fourth vertex* (*Theorem* 3.3).

The theorem above has a generalization which concerns a complete set of reduction rules. A set of reduction rules will be called *complete* if they are sufficient to reduce all and only graphs from a given class to a given canonical form. In this context, a reduction rule is safe if it cannot take a member of the class outside the class.

THEOREM 3.5. *Each of the reductions from a complete set of confluent rules is safe.*

Proof. Let us assume, to the contrary, that a graph x in a class C has a successor (reduct) y not in C. Since the reduction rules are complete, x can be reduced to the canonical form z, while y can not. However, confluence implies that all reducts of x have a common successor. □

The results of this section yield an $O(n^3)$ algorithm for finding an embedding 3-tree of a graph with n vertices or deciding that no such embedding exists. The time required for performing the S_3' reductions once the vertex order has been decided is, with suitable data structures (see Wald and Colbourn [14]) $O(n)$. In order to find the next vertex to reduce, first in time $O(n)$ select a vertex of degree $\leqq 2$ or a vertex of degree 3 with either (i) two adjacent neighbors (in which case the selected vertex is reduced) or (ii) three neighbors of degree 3 with overlapping neighborhoods so that the selected vertex is x in a subgraph isomorphic to C' of Fig. 3 (in this case the neighbors of x are reduced). If no such vertex exists, check for all $O(n^2)$ degree 3

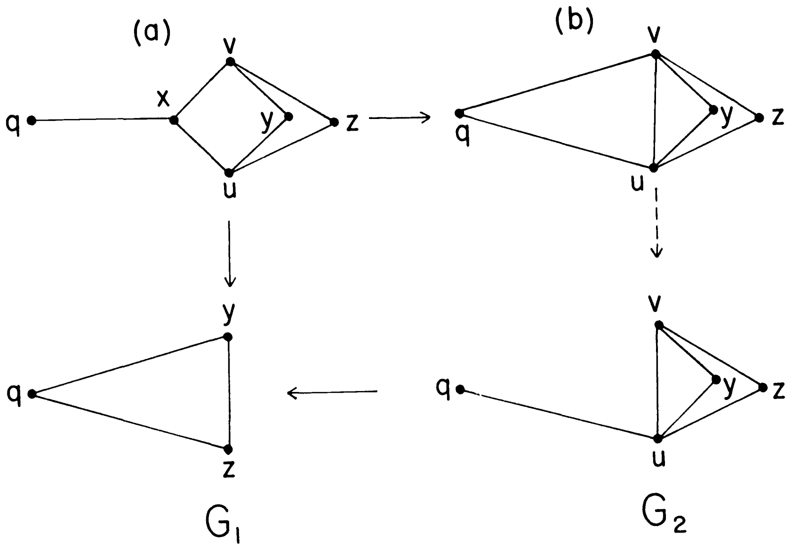


FIG. 7. *Reduction of a partial* 3*-tree G, case* (iii).

vertex pairs if they have common neighborhood (in which case they are reduced). Each pair can be processed in constant time. The total worst-case time for finding the reduction order or deciding that no such order exists is clearly $O(n^3)$. Our referee has pointed out that this can be considerably improved by two modifications: The first consists in having all neighborhoods of degree 3 vertices (where vertices have been numbered in an arbitrary order and each neighborhood is represented as an ordered triple) as keys in a structure supporting insert, search and delete in time $O(\log n)$ (e.g. an AVL tree). This makes it unnecessary to examine all pairs in order to find configuration C''. The second improvement consists in having a list of ready vertices (that fulfill the conditions for safe reduction). Each reduction made implies a neighborhood change for at most 3 vertices and removal of at most 4 vertices from the ready list (which means that the total number of additions and deletions in the ready list is $O(n)$). With these improvements worst-case processing times of $O(n \log n)$ seem possible.

4. Conclusion and further research. We have found a set of confluent reductions on graphs such that any graph can be reduced to the empty graph if and only if it is a partial 3-tree. This set of reductions yields a polynomial time algorithm for recognizing partial 3-trees and embedding them in full 3-trees. This generalizes the previously known recognition algorithm for partial 2-trees of Wald and Colbourn's [14].

Already for the case of $k = 4$, there is no easy generalization of our methods used in recognition of partial 3-trees. A solution to this problem for arbitrary k would have significant practical applications, since graph algorithms based on decomposition are frequently used, even though only heuristic decomposition strategies are known. The cost of such a decomposition algorithm is often exponential in the order of the articulation sets used. Thus, the minimax solution given by a k-tree embedding for the minimum value of k is clearly highly relevant.

In a preliminary presentation of this research we describe families of safe (but not necessarily complete) reductions for general partial k-trees [2, Thms. 4.1–4.6]. We have programmed these reduction rules and tested them on partial k-trees generated by Monte-Carlo techniques. For small values of k (up to 7), almost all of the graphs were correctly recognized, but the failure rate grew with increasing k. The existing, incomplete set of safe reduction rules could thus be used as another heuristic decomposition method. It differs from most such methods in its "bottom-up" (rather than "top-down") approach. This method has the worst case complexity of order $O(n^k)$, which compares favorably with the $O(n^{k+2})$ complexity of the only known complete recognition algorithm for partial k-trees [3]. If complete sets of safe reductions are found for arbitrary k and if the improvements suggested by the referee (see § 4) carry over to this case, one could even expect an algorithm for recognizing partial k-trees in time $O(f(k)n \log n)$. Here $f(k)$ is probably exponential since the general recognition problem for partial k-trees (with the value of k given in the problem instance) is NP-complete [3].

Acknowledgments. The referee has suggested significant improvements and corrections. This research was supported by the National Science Foundation under grant INT-8318441 and by the Swedish Board for Technical Development under grant 83-3719.

REFERENCES

[1] S. Arnborg, *On the complexity of multivariable query evaluation*, FOA Rapport C20292-D8, National Defence Research Institute, Stockholm, Sweden, 1979.

[2] S. Arnborg and A. Proskurowski, *Characterization and recognition of partial k-trees*, TRITA-NA 8402, Royal Institute of Technology, Sweden, 1984.

[3] S. Arnborg, D. G. Corneil and A. Proskurowski, *Complexity of finding embeddings in a k-tree*, TRITA-NA 8407, Royal Institute of Technology, Sweden, 1984.

[4] L. W. Beineke and R. E. Pippert, *Properties and characterizations of k-trees*, Mathematika, 18 (1971), pp. 141-151.

[5] R. J. Duffin, *Topology of series-parallel networks*, J. Math. Anal. Appl., 10 (1965), pp. 303-318.

[6] A. M. Farley, *Networks immune to isolated failures*, Networks, 11 (1981), pp. 255-268.

[7] A. M. Farley and A. Proskurowski, *Networks immune to isolated line failures*, Networks, 12 (1982), pp. 393-403.

[8] G. Huet and D. Oppen, *Equations and rewrite rules: a survey*, in Formal Languages: Perspective and Open Problems, R. Book, ed., Academic Press, New York, 1980.

[9] P. C. Liu and R. C. Geldmacher, *Graph reducibility*, in Proc. Seventh S-E Conf. Combinatorics, Graph Theory, and Computing, Utilitas Mathematica, Winnipeg, 1976, pp. 433-455.

[10] E. M. Neufeld and C. J. Colbourn, *The most reliable series-parallel networks*, TR 83-7, Dept. Comp. Sci., Univ. Saskatchewan, 1983.

[11] D. Rose, *Triangulated graphs and the elimination process*, J. Math. Anal. Appl., 32 (1970), pp. 597-609.

[12] ———, *On simple characterization of k-trees*, Discrete Math., 7 (1974), pp. 317-322.

[13] D. Rose, R. E. Tarjan and G. S. Lueker, *Algorithmic aspects of vertex elimination on graphs*, SIAM J. Comput., 5 (1976), pp. 266-283.

[14] J. A. Wald and C. J. Colbourn, *Steiner trees, partial 2-trees, and minimum IFI networks*, Networks, 13 (1983), pp. 159-167.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

015

AN APPLICATION OF THE SINGULAR VALUE DECOMPOSITION TO MANIPULABILITY AND SENSITIVITY OF INDUSTRIAL ROBOTS*

MASAKI TOGAI†

Abstract. In designing and evaluating industrial robots, it is important to find optimal configurations and locate optimum points in the workspace for the anticipated tasks. In the current paper the singular value decomposition and perturbation analysis are applied to the Jacobian of robot kinematics; the condition number of the Jacobian is then proposed to be a measure of the "nearness" to degeneracy. Then qualitative measures called kinematic "manipulability" and "sensitivity" are proposed. Some properties of proposed measures are investigated and the relation between these measures are discussed. Optimal postures of various types of industrial robots are obtained.

AMS(MOS) subject classifications. 15, 65

1. Introduction. In designing and evaluating industrial robots, it is important to find optimum configurations, or postures, and locate optimum points in the workspace for the anticipated tasks. This becomes increasingly important in high precision assembly. Several measures of workspace are available. The size of reachable volume is an important performance measure [1]. To obtain full mobility throughout its range of motion, the ideal manipulator would have no singularities, or degeneracies, in its workspace. In general, a manipulator becomes degenerate in its workspace; therefore, the "nearness" to the degeneracy is also an important measure. Yoshikawa [2], [3] calls it "manipulability." Another important measure of workspace quality is the accuracy with which the task would be achieved. Particularly if the magnitude of accuracy of the manipulator is comparable to that of the anticipated tasks such as high precision assembly tasks, this measure is extremely important.

In this paper a new qualitative measure for manipulability is proposed. Advantages of the proposed measure over Yoshikawa's definition are discussed. Another qualitative measure of a manipulator's ability of accurately positioning and orienting a manipulator, so-called *sensitivity*, is proposed. Some properties of proposed manipulability and sensitivity are investigated and the relation between these measures are discussed. Optimal postures of various types of manipulators are obtained. Some computational consideration of proposed manipulability and sensitivity are also discussed.

2. Manipulability: A new definition. Yoshikawa [2] proposed $w=\sqrt{\det(JJ^T)}$ for a qualitative measure of manipulating ability of robot arms, and called it *manipulability*. According to the singular valued decomposition (SVD) theorem [4], assuming J is m-by-n matrix there exist orthogonal matrices $U\in R^{m\times m}$ and $V\in R^{n\times n}$ such that

$$J=U\sum V^T, \tag{1}$$

where

$$\sum=\left[\begin{array}{cccc|c} \sigma_1 & & 0 & & \\ & \sigma_2 & & & 0 \\ 0 & & \ddots & & \\ & & & \sigma_m & \end{array}\right]\in R^{m\times n}$$

* Received by the editors January 15, 1985. This paper was presented at Second SIAM Conference on Applied Linear Algebra, Raleigh, North Carolina, April 29–May 2, 1985.

† AT&T Bell Laboratories, Holmdel, New Jersey 07733.

with

$$\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_m \geqq 0.$$

The entries σ_i in $\sum$ are called the *singular values* of J. Therefore, according to Yoshikawa's definition, the manipulability is given by the product of the singular values $\sigma_1, \sigma_2, \cdots, \sigma_m$, i.e.

$$w = \sigma_1 \sigma_2 \cdots \sigma_m.$$

Mathematically the manipulability is a measure of the "nearness" of J to degeneracy. For this purpose, the determinant of JJ^T is a "terrible measure" of nearness, because it depends on not only the size of J, m, but also the scaling factor such as the product of all singular values [5].

In this paper the *condition number* of J, or cond(J), is proposed to be a better measure of the manipulability.

Definition (*manipulability*). The *manipulability* M of a robot arm is defined by the condition number of its Jacobian:

$$M = \operatorname{cond}(J) = \|J\| \, \|J^{-1}\|, \tag{2}$$

where $\|\cdot\|$ is a norm.

Note that mathematically the condition number is a much more precise and reliable measure of "nearness" to singularity than quantities such as the determinant or the smallest pivot [4]. If $\|\cdot\|$ is a Eucledian norm, then $M = \sigma_{\max}/\sigma_{\min}$, where $\sigma_{\max}$ and $\sigma_{\min}$ are the largest and the smallest absolute values of singular values of J, respectively; therefore, the measure is not affected either by the size of J or by the scaling factor. Another advantage of this measure is that the condition number is greatest when $\sigma_{\min}$ is close to zero, so that the near-degeneracy configuration is the most sensitive. Points in the workspace that minimize the condition number of the Jacobian, i.e. manipulability, are the best conditional ones to operate manipulators. The best conditioning possible in terms of manipulability occurs when $M = \operatorname{cond}(J) = 1$. Such best conditional points, called *isotropic points* [6], may or may not exist for a given design. For example, a Cartesian type manipulator would have manipulability of one everywhere within the workspace because $M = \operatorname{cond}(J) = \operatorname{cond}(\mathbf{I}) = 1$, where $\mathbf{I}$ is a unit matrix.

3. Kinematic sensitivity functions. The location of most accurate operational points in the workspace is a useful design and evaluation consideration, especially for precise assembly tasks. The effect of the parameter joints on the positional accuracy in the workspace can be expressed in terms of *sensitivity.*

Definition (*kinematic sensitivity matrix*). The parameter-induced trajectory deviation [7] is expressed by

$$\Delta\mathbf{X} = S(\theta_0)\Delta\theta \tag{3}$$

where $S(\theta_0)$ is called the *kinematic sensitivity matrix.*

Note that (3) is equivalent to the familiar matrix form $\Delta\mathbf{X} = J(\theta_0)\Delta\theta$ where $J(\theta_0)$ is the configuration dependent Jacobian matrix. In other words, a Jacobian matrix can be considered as a kinematic sensitivity matrix of a given robot manipulator. Therefore, (3) can be also written in the familiar matrix form such as:

$$\Delta\mathbf{X} = J(\theta_0)\Delta\theta. \tag{4}$$

The instantaneous velocity in the task space is expressed in terms of the kinematic sensitivity and the joint velocity θ:

$$\dot{X} \approx (\Delta X/\Delta T) = S(\theta_0)(\Delta\theta/\Delta T) = S(\theta_0)\dot{\theta}. \tag{5}$$

The velocity of a manipulator in the workspace is defined by the kinematic sensitivity, i.e. the Jacobian, and the speed of each joint. In other words, the maximum controllable velocity of a manipulator in the workspace is decided by the kinematic sensitivity, i.e. the Jacobian, and the maximum controllable velocity of each joint.

The effect of the velocity change in joint coordinates on the velocity change in the workspace can also be expressed in terms of the sensitivity function.

DEFINITION (*kinematic velocity sensitivity*). The *kinematic velocity sensitivity* $S_{\dot\theta}^{\dot X}$ is a measure of the manipulator's velocity response in the workspace to velocity variations in the joint coordinates and is given by

$$\begin{aligned} S_{\dot\theta}^{\dot X} &= \frac{\text{(relative velocity change in workspace)}}{\text{(relative velocity change in joint coordinates)}} \\ &= \frac{\|d\dot x\|/\|\dot x\|}{\|d\dot\theta\|/\|\dot\theta\|} \end{aligned} \tag{6}$$

where

$\dot X$ and $d\dot X$: velocity and velocity error in the workspace,

$\dot\theta$ and $d\dot\theta$: velocity and velocity error in the joint coordinates,

$\|\cdot\|$: norm.

Applying linear system error analysis [5] to (5), the relative velocity errors in the joint coordinates are bounded by the product of the cond (S), or cond (J), and the relative velocity error in the workspace.

$$\frac{\|d\dot\theta\|}{\|\dot\theta\|} \leqq \operatorname{cond}(J)\frac{\|d\dot X\|}{\|\dot X\|}. \tag{7}$$

In other words, the condition number of J, or manipulability M, gives a measure of how much error in $\dot X$ may be magnified in errors in $\dot\theta$. From (7) it can be easily shown that an inverse value of the manipulability M, i.e. $1/M$, gives the lower bound of the kinematic velocity sensitivity $S_{\dot\theta}^{\dot X}$:

$$S_{\dot\theta}^{\dot X} = \frac{\|d\dot X\|/\|\dot X\|}{\|d\dot\theta\|/\|\dot\theta\|} \geqq \frac{1}{M} \tag{8}$$

where $M = \operatorname{cond}(J) = \|J\|\,\|J^{-1}\|$. Since $M \geqq 1$, $S_{\dot\theta}^{\dot X}$ and $1/M$ are bounded as follows:

$$1 \geqq S_{\dot\theta}^{\dot X} \geqq \frac{1}{M} \geqq 0, \tag{9}$$

it can be concluded from (9) that the best conditioning possible points in the workspace in terms of manipulability are the worst conditioning possible points in terms of velocity sensitivity, and vice versa. Thus, the relation between manipulability and velocity sensitivity is established.

4. Manipulability and sensitivity of various types of manipulators. The simulations on manipulability and sensitivity of various types of manipulators demonstrate the utility and effectiveness of the proposed measures.

The optimal postures of a two-joint link in terms of manipulability and sensitivity with various second link lengths are shown in Fig. 1 and Fig. 2, respectively. It is worthwhile to note that the optimal manipulatable postures of the first link remain the same for various lengths of second link as long as 0.707 $l_1 \leqq l_2 \leqq l_1$. This is similar to

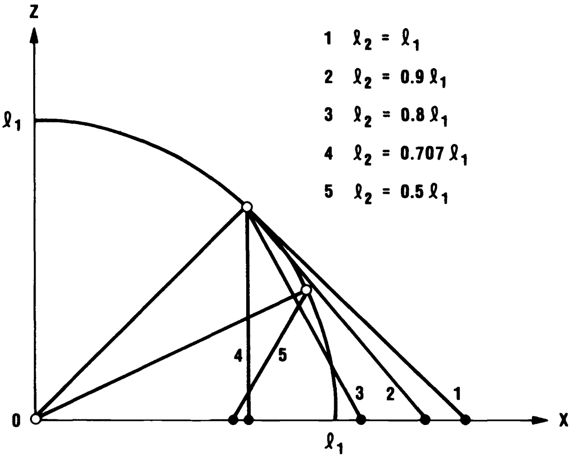


FIG. 1. *Postures with optimal manipulability.*

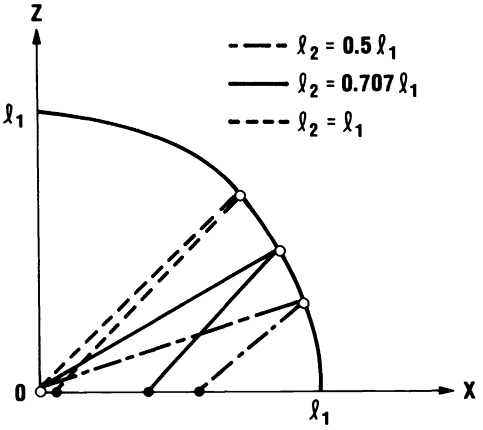


FIG. 2. *Optimal postures for precise assembly.*

our arms: in general the upper arm remains almost in the same posture as long as the lower arm and hand can reach the anticipated points in the workspace.

Salisbury and Craig [6] used the condition number of J^T to find optimal posture of fingers to minimize error propagation from input torque to output forces because the torque τ and the force F are related by the equation $\tau = J^T F$. Yoshikawa [2], [3] proposed $w = \sqrt{\det (JJ^T)}$ for a measure of manipulability as mentioned previously. In Fig. 3, the optimal postures of a two-joint link in terms of these measures are depicted along with the optimal postures determined by the measure proposed in this paper. While optimal postures determined by Yoshikawa's measure give the second joint angles in the vicinity of $\pi/2$ for various length of the second link, optimal postures determined by the proposed measure in this paper give the first link in the same posture as long as it can. Similar optimal postures are obtained for other articulated manipulators which contain two-joint links.

Note that the optimal postures determined by the measure proposed by Yoshikawa for cylindrical and polar type manipulators are given when the prismatic joint at the arm is stretched out and the tip of the arm reaches the outer boundary of the work

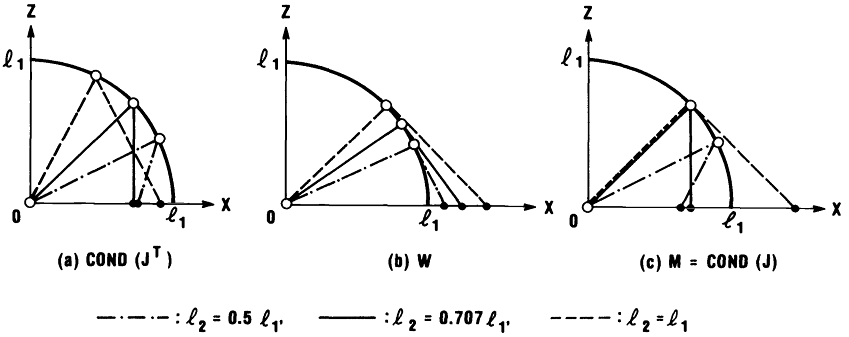


FIG. 3. *Comparison of optimal postures based on various definitions.*

space. As Yoshikawa pointed out in his paper [3], this is very inconvenient. This inconvenience is caused by the "scaling factor" which was mentioned previously in this paper. The proposed measure is free from this scaling factor. Therefore, the proposed measure gives the optimal postures when the prismatic joint is pulled in and the tip of arm reaches the inner boundary of the work space. These are more reasonable postures than those determined by Yoshikawa's definition. Figure 4 shows the manipulability M of a cylindrical manipulator as a function of the sliding distance of a prismatic joint of the arm. It clearly shows that the optimal posture occurs when the tip of the arm reaches the inner boundary of the workspace. Note that the position in the direction of Z-axis does not affect the manipulability.

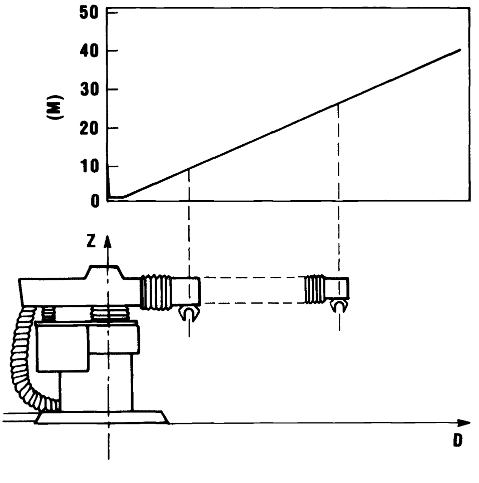


FIG. 4. *Manipulability of a cylindrical manipulator.*

5. Summary. A reliable and useful measure of robot manipulability has been proposed; kinematic sensitivities has been defined. It has also been proved that the velocity sensitivity is bounded by the inverse of the measure of manipulability. In other words, the best conditioning possible postures of robot manipulator in terms of manipulability are the worst possible postures in terms of velocity sensitivity.

Optimal postures in terms of kinematic manipulability and sensivitity of various types of manipulators are obtained. It has been demonstrated that the measures proposed in this paper to find optimal postures of arms are more suitable than other measures.

Appendix. Numerical calculation of the condition number. The actual computation of cond (J) involves knowing J^{-1}. If s_j are the columns of J and $\bar{s}_j$ are the columns of J^{-1} then in terms of the vector norm we are using

$$M = \text{cond}\,(J) = \max_j \|s_j\| \cdot \max_j \|\bar{s}_j\|. \tag{A1}$$

It is easy to compute $\|J\|$, but finding $\|J^{-1}\|$ would roughly triple the time required for Gaussian elimination. Fortunately, the exact value of cond (J) is rarely required. Any reasonably good estimate of it is satisfactory. In this paper, the manipulability is estimated by

$$M = \text{cond}\,(J) \approx \max_j \|s_j\| \frac{\|Z\|}{\|Y\|}, \tag{A2}$$

where Y and Z are two vectors such that $\|Z\|/\|Y\| \approx \|J^{-1}\|$. This involves solving two systems of equations

$$J^T Y = E, \qquad JZ = Y,$$

where J^T is the transpose of J and E is a components ±1 chosen to maximize the growth during the back substitution for [8].

Two subroutines used to obtain the condition of J are decomp ( ) and solve ( ). Decomp ( ) carries out the part of Gaussian elimination and saves the multipliers and the pivot information. Solve ( ) uses these results to obtain the solution. Decomp ( ) also returns an estimate of the condition of the matrix J. Such an estimate is a much more reliable and useful measure of nearness to singularity than quantities such as the determinant or the smallest pivot. Decomp ( ) can be used to compute determinants. The last component of the pivot vector returns +1 if an even number of row interchanges is used, and the value −1 if an odd number is used. This value is multipled by the product of the diagonal elements of the output matrix to obtain the determinant. Both subroutines are coded in C and implemented on VAX 11/780.

REFERENCES

[1] K. SUGIMOTO AND J. DUFFY, *Determination of extreme distances of a robot hand—Part* 1, *a general theory*, ASME J. Mech. Design, 103 (1981), pp. 631-636.

[2] T. YOSHIKAWA, *Analysis and control of robot manipulators with redundancy*, Preprints of the 1st International Symposium of Robotics Research, Aug. 28-Sept. 2, 1983.

[3] T. YOSHIKAWA, *Measure of manipulability for robot manipulators*, J. Robotic Society of Japan, 2, 1 (1984), pp. 63-67. (In Japanese.)

[4] V. C. KLEIN AND A. J. LAUB, *The singular value decompostion: its computation and some applications*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 164-176.

[5] G. STRANG, *Linear Algebra and Its Applications*, Academic Press, New York, 1976.

[6] J. K. SALISBURY AND J. J. CRAIG, *Articulated hands: force control and kinematic issues*, Internat. J. Robotics Research, 1 (1982), pp. 4-17.

[7] P. M. FRANK, *Introduction to System Sensitivity Theory*, Academic Press, New York, 1978.

[8] G. E. FORSYTHE, M. A. MALCOLM AND C. B. MOLER, *Computer Methods for Mathematical Computations*, Prentice-Hill, Englewood Cliffs, NJ, 1977.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

016

A MULTIPLIER METHOD FOR IDENTIFYING KEYBLOCKS IN EXCAVATIONS THROUGH JOINTED ROCK*

J. L. DELPORT† AND D. H. MARTIN†

Abstract. G. H. Shi and R. E. Goodman have recently pointed out the practical importance of the keyblock principle in supporting underground and surface excavations in jointed rock and have given an elegant procedure involving stereographic projection and graphics for identifying the shapes of keyblocks. This paper presents the mathematical analysis of an entirely different keyblock characterisation and identification algorithm. The characterisation rests upon Tucker's Theorem of the Alternative, and the algorithm requires only the execution of a few linear programming type pivot operations and sign tests.

Key words. keyblock, keystone, stability of excavations, theorem of the alternative

AMS(MOS) subject classification. 49D33

1. Introduction. Following Shi and Goodman [5], [6], we consider a body of rock that is traversed by a number of families of more or less parallel discontinuity planes or faults. The rock body is divided by these discontinuities into blocks and slabs and, when excavating a tunnel, some of these may be liable to loosen and fall into the tunnel. As described by Shi and Goodman, once these first blocks have fallen, others, which were previously held in place by them, may also be free to all, possibly leading progressively to a major collapse. Shi and Goodman have termed these critical blocks that must fall first, *keyblocks*, and have shown how they may be identified using a graphical method involving stereographic projection. See also [1].

In this paper we develop an efficient, purely computational algorithm for identifying the keyblocks. Execution of the algorithm requires only the reduction of a system of three linear equations, followed by successive *LP*-type pivot operations and tests for the signs of the entries. In practice, the number of families of discontinuity planes seldom exceeds five, and the algorithm can easily be executed in such cases on a programmable pocket calculator.

This paper deals with what may be termed the convex case, in which the excavation is taken to be convex and bounded by finitely many planes, and the visible blocks of rock, which may or may not be keyblocks, are convex apart from nonconvexities introduced by the excavation itself. The complicating effects of other nonconvexities will be dealt with in a subsequent work.

2. Discontinuity planes and rock blocks. Suppose there are $m \geqq 3$ families of discontinuity planes. Each family is idealised to a family of parallel planes, and hence all members of the ith family have a common unit normal vector $\mathbf{n}_i$ and are given by linear equations of the form

$$\mathbf{n}_i \cdot \mathbf{r} = c_i \tag{1}$$

for different values of the constant c_i. Here $\mathbf{r} = (x, y, z)$ is the position vector of a representative point on the plane, and the dot denotes the scalar product. We adopt the convention that the normal vectors $\mathbf{n}_i$ are oriented upwards. Then points above the plane (1) are given by

$$\mathbf{n}_i \cdot \mathbf{r} \geqq c_i,$$

* Received by the editors August 16, 1984.

† National Research Institute for Mathematical Sciences of the CSIR, PO Box 395, Pretoria 0001, South Africa.

while those below the plane (1) are given by

$$\mathbf{n}_i \cdot \mathbf{r} \leqq c_i.$$

Equivalently, if we introduce the orientation scalar

$$e_i = \begin{cases} 1 & \text{for upward orientation,} \\ -1 & \text{for downward orientation,} \end{cases}$$

then the inequality

$$e_i \mathbf{n}_i \cdot \mathbf{r} \geqq e_i c_i$$

gives points above or below the plane (1) according to the choice of orientation. *We make a standing assumption that no trio of the normals* $\mathbf{n}_i$ *is coplanar.*

The discontinuity planes divide the body of rock into blocks, and the shape of a particular block depends upon above which and below which planes it is, while its size is influenced by the relative position of the planes. A *convex* block bounded by m different planes is thus given by simultaneous inequalities

$$(2) \qquad e_i \mathbf{n}_i \cdot \mathbf{r} \geqq e_i c_i, \qquad i = 1, 2, \cdots, m,$$

and its shape is thus characterised by an orientation pattern $(e_1, e_2, \cdots, e_m)$. It remains to introduce the excavation that exposes the block. For simplicity we suppose that only one plane face of the excavation intersects the block—this assumption is relaxed later. Let $\mathbf{N}$ denote unit normal to this excavation face, oriented to point out of the rock face into the excavation. Then the rock block is visible in the excavation face

$$\mathbf{N} \cdot \mathbf{r} = d$$

and is given by the inequalities

$$(3) \qquad \begin{aligned} &e_i \mathbf{n}_i \cdot \mathbf{r} \geqq e_i c_i, \qquad i = 1, 2, \cdots, m, \\ &\mathbf{N} \cdot \mathbf{r} \leqq d. \end{aligned}$$

We make the further standing assumption that $\mathbf{N}$ is not coplanar with any pair of the normals $\mathbf{n}_1, \cdots, \mathbf{n}_m$. This is equivalent to the assumption that discontinuity planes from different families show as nonparallel lines in the excavation face.

3. Keyblocks. The characteristic properties of a keyblock given by the system (3) are first, that it is of finite size—i.e., *constitutes a bounded point set*—and second, that it is *movable without any neighboring block having to move simultaneously.* A standard theorem about convex sets (see, for example, [4, Thm. 8.4, p. 64]) ensures that the set (3) is bounded if and only if the associated homogeneous system

$$(4) \qquad \begin{aligned} &e_i \mathbf{n}_i \cdot \mathbf{r} \geqq 0, \qquad i = 1, 2, \cdots, m, \\ &\mathbf{N} \cdot \mathbf{r} \leqq 0 \end{aligned}$$

has no nontrivial solution.

The movability of the block requires the existence of a nonzero translation vector $\mathbf{v}$ that will not entail the block encroaching upon any adjoining block, i.e. such that

$$(5) \qquad \mathbf{v} \neq \mathbf{0}, \qquad e_i \mathbf{n}_i \cdot \mathbf{v} \geqq 0, \quad i = 1, 2, \cdots, m.$$

The question treated in this paper is the following: Which among the 2^m different orientation patterns $(e_1, e_2, \cdots, e_m)$ correspond to keyblocks—i.e. which are such that (4) has no nontrivial solution, while (5) does have? Shi and Goodman have noted that the number of different keyblock orientation patterns is

$$\binom{m-1}{2}=\frac{1}{2}(m-1)(m-2). \tag{6}$$

The characterisation (4), (5) differs notationally but agrees mathematically with that used by Shi and Goodman in the cited references. However, from this point on our approach to the identification of those orientation patterns $(e_1, e_2, \cdots, e_m)$ that correspond to potential keyblocks is quite different to that of Shi and Goodman. Whereas their technique hinges on a stereographic projection of possible solutions ($\mathbf{r}$ or $\mathbf{v}$) of (4) and (5) onto a plane, ours involves multipliers that are introduced using a so-called theorem of the alternative and leads to a simple algorithm that requires only the solution of systems of linear equations and sign tests.

4. Multiplier characterisation of keyblocks. We first show that since no trio of normals $\mathbf{n}_i$ is coplanar, the existence of a solution $\mathbf{v}$ such that (5) holds is equivalent to the existence of a solution $\mathbf{v}$ to the ostensibly stronger inequalities

$$e_i\mathbf{n}_i \cdot \mathbf{v} > 0, \qquad i = 1, 2, \cdots, m. \tag{7}$$

Clearly, if (7) is consistent, so is (5). For the converse suppose (7) is inconsistent, but (5) consistent. By Tucker's Theorem of the Alternative[1] the inconsistency of (7) implies the existence of nonnegative multipliers $\mu_1, \cdots, \mu_m$ such that

$$\sum_{i=1}^{m} \mu_i e_i \mathbf{n}_i = \mathbf{0}, \quad \text{each } \mu_i \geqq 0, \quad \text{not all } \mu_i \text{ zero}. \tag{8}$$

But if $\mathbf{v}$ is any solution of the system (5), then from (8) we have

$$0 = \sum_{i=1}^{m} \mu_i(e_i\mathbf{n}_i \cdot \mathbf{v}),$$

where each term is nonnegative. Hence for every index i, either $\mu_i = 0$ or $e_i\mathbf{n}_i \cdot \mathbf{v} = 0$. Since at most two of the normals can be orthogonal to $\mathbf{v}$, at most two of the multipliers in (8) are nonzero, leaving at most two terms. But this means two of the normals are parallel, which is false. Thus the consistency of (5) is equivalent to that of (7) and, by Tucker's Theorem of the Alternative, is equivalent to the *inconsistency* of the system (8).

Turning to the system (4), since the normals $\mathbf{n}_1, \cdots, \mathbf{n}_m$ are not coplanar, a solution $\mathbf{r}$ will be nontrivial if and only if not all of the scalar products $e_i\mathbf{n}_i \cdot \mathbf{r}$ vanish. Hence the nonexistence of a nontrivial solution to (4) is equivalent to the system

$$\begin{aligned} &e_i\mathbf{n}_i \cdot \mathbf{r} \geqq 0 \quad \text{for all } i, \quad \text{and not all zero,} \\ &\mathbf{N} \cdot \mathbf{r} \leqq 0 \end{aligned} \tag{9}$$

being inconsistent. By Tucker's Theorem of the Alternative this inconsistency is equivalent to the consistency of the system

$$-\mu\mathbf{N} + \sum_{i=1}^{m} \mu_i e_i \mathbf{n}_i = \mathbf{0}, \qquad \mu \geqq 0, \quad \text{all } \mu_i > 0. \tag{10}$$

[1] See, for example [3, p. 29]. For the possible convenience of readers, this theorem is stated in an Appendix hereto. It is applied here with the $m \times m$ identity matrix in the role of the matrix B and the $3 \times m$ matrix with columns given by the vectors $e_i\mathbf{n}_i$, $i = 1, 2, \cdots, m$ in the role of D, C being vacuous.

Thus an orientation pattern $(e_1, \cdots, e_m)$ corresponds to a keyblock if and only if the system (10) is consistent while (8) is inconsistent.

Clearly we cannot have (8) being inconsistent while (10) is consistent with $\mu = 0$. Hence it suffices to consider (10) with $\mu = 1$. But then the inconsistency of (8) is precisely the condition that the solution set of (10), if nonempty, should be bounded. Setting $e_i\mu_i = \lambda_i$ in (10), this establishes the following characterisation.

A visible convex block (3) *is a keyblock if and only if its orientation pattern* $(e_1, \cdots, e_m)$ *is such that the set of multiplier m-tuples* $(\lambda_1, \cdots, \lambda_m)$ *that satisfy the equation*

$$\sum_{i=1}^{m} \lambda_i \mathbf{n}_i = \mathbf{N} \tag{11}$$

and have the sign pattern

$$\operatorname{sgn} \lambda_i = e_i, \qquad i = 1, 2, \cdots, m \tag{12}$$

is both nonempty and bounded.

The advantage of this characterisation is that it involves a fixed system (11) of equations that does not depend on the choice of orientation pattern. Geometrically, (11) defines a plane Π of $m-3$ dimensions in the m-dimensional space $\mathbb{R}^m$ with coordinates $(\lambda_1, \cdots, \lambda_m)$. In $\mathbb{R}^m$, each orientation pattern $(e_1, \cdots, e_m)$ corresponds via (12) to a particular orthant in $\mathbb{R}^m$, and keyblock orientations correspond to those orthants that have a nonempty but bounded intersection with the plane Π.

5. Determination of the orientation patterns of keyblocks. In $\mathbb{R}^m$, each coordinate hyperplane $\lambda_k = 0$ intersects the plane Π of all solutions of (11) in a (relative) hyperplane that partitions Π into two half-planes, each characterised by the sign of λ_k. Let $(e_1, e_2, \cdots, e_m)$ be the orientation pattern of a keyblock. This will be so iff the simultaneous inequalities

$$e_k\lambda_k \geqq 0, \qquad k = 1, 2, \cdots, m, \tag{13}$$

together with (11), define a nonempty compact polyhedron P on Π with nonempty interior relative to Π. At vertices of this polyhedron $m-3$ of the coordinates λ_k vanish. Because of the standing assumption that $\mathbf{N}$ is not coplanar with any pair of the normals $\mathbf{n}_k$, it is clear from (11) that at any vertex precisely three of the coordinates λ_k are nonzero. The vertices are therefore distinct, and may be labelled according to the nonzero coordinates: V_{ijk} denotes that point of Π at which $\lambda_h = 0$ for all $h \neq i, j, k$—here i, j, k will always signify distinct indices.

For any fixed index k, the quantity $e_k\lambda_k$, regarded as an affine function on Π, is positive throughout the relative interior of our compact polyhedron P, and hence achieves a positive maximum on P—i.e., subject to the constraints (11) and (13). By a fundamental theorem of linear programming (see, for example, [2]) this maximum is achieved at a basic feasible solution—i.e., at some vertex of P. Since $e_k\lambda_k > 0$ at this vertex, it is necessarily of the form V_{kij} for some indices i, j, and the variables λ_k, λ_i, λ_j are the only nonzero ones at V_{kij}. Since $\mathbf{n}_k$, $\mathbf{n}_i$ and $\mathbf{n}_j$ are not coplanar, equations (11) may be solved for λ_k, λ_i and λ_j in terms of the other variables to produce an equivalent reduced system of the form

$$\begin{aligned} \lambda_k + \sum_{h \neq i,j,k} a_{kh}\lambda_h &= b_k, \\ \lambda_i + \sum_{h \neq i,j,k} a_{ih}\lambda_h &= b_i, \\ \lambda_j + \sum_{h \neq i,j,k} a_{jh}\lambda_h &= b_j, \end{aligned} \tag{14}$$

as in the well-known simplex method of linear programming. Here b_k, b_i, b_j are the nonzero values of λ_k, λ_i and λ_j respectively at the maximum, and it follows immediately that

$$(15)\qquad e_k = \operatorname{sgn} b_k, \quad e_i = \operatorname{sgn} b_i, \quad e_j = \operatorname{sgn} b_j.$$

We may rewrite the first equation in (14) as

$$e_k\lambda_k + \sum_{h \neq i,j,k} e_k e_h a_{kh}(e_h\lambda_h) = e_k b_k,$$

and then, from the fact that the function $e_k\lambda_k$ achieves its maximum subject to (14) (which is equivalent to (11)) and (13), it follows that we must have

$$e_k e_h a_{kh} \geqq 0 \quad \text{for all } h \neq k, i, j.$$

It will be shown at the end of this section that none of the coefficients a_{kh}, a_{ih}, a_{jh} appearing in (14) can be zero. Accepting this for the moment, it now follows that the maximizing vertex V_{kij} is unique and that

$$(16)\qquad e_h = e_k \operatorname{sgn} a_{kh}, \qquad h \neq k, i, j.$$

To summarise, this proves that for any index k held fixed, to each keyblock orientation pattern $(e_1, \cdots, e_m)$, there corresponds a unique index pair $i, j \neq k$ such that the orientation pattern is generated from (14) by the formulas (15) and (16).

We now establish a converse statement. Still holding the index k fixed, let i, j be any pair of distinct indices different from k, and consider the *simplex* S_{ij} in Π generated by the $m-2$ vertices

$$V_{hij}, \ h \neq i, j.$$

For any $h \neq i, j$, the face of S_{ij} opposite vertex V_{hij} lies in the hyperplane $\lambda_h = 0$, whereas at V_{hij}, the coordinate λ_h is nonzero. It follows that throughout the relative interior of S_{ij}, the coordinate λ_h has the same sign as it has at V_{hij}. In equations (14) (which are equivalent to (11)) we see that for $h \neq k, i, j$, if we set all coordinates zero other than λ_h, λ_i, λ_j, the first equation reduces to

$$a_{kh}\lambda_h = b_k,$$

from which it follows that the sign of λ_h at V_{hij} is given by

$$(17)\qquad \operatorname{sgn} \lambda_h = \operatorname{sgn}(b_k/a_{kh}) = (\operatorname{sgn} b_k)(\operatorname{sgn} a_{kh}), \qquad h \neq k, i, j,$$

For $h = k$, it follows directly from (14) that at V_{kij},

$$(18)\qquad \operatorname{sgn} \lambda_k = \operatorname{sgn} b_k.$$

Thus (17) and (18) hold throughout the simplex S_{ij}.

The situation for λ_i and λ_j is different—they may or may not change sign within S_{ij}. However, from (14), at the vertex V_{kij} of S_{ij} they have the signs

$$(19)\qquad \operatorname{sgn} \lambda_i = \operatorname{sgn} b_i, \qquad \operatorname{sgn} \lambda_i = \operatorname{sgn} b_j.$$

It follows that points in the relative interior of S_{ij} sufficiently close to the vertex V_{kij} have the sign pattern given by (17), (18) and (19). Furthermore, since at least one sign changes across any boundary face of the simplex S_{ij}, this pattern of signs occurs only at points of a bounded subset of Π. It therefore corresponds to a keyblock orientation pattern. It remains only to note that (17), (18) and (19) can be written as $\operatorname{sgn} \lambda_h = e_h$ for all h with $(e_1, \cdots, e_m)$ given by (15) and (16). This proves the following result.

For each index k held fixed, there is a one-to-one correspondence between keyblock orientation patterns $(e_1, \cdots, e_m)$ *and pairs of distinct indices i, j different from k such that the orientation pattern is generated by reducing the equations* (11) *to the form* (14) *and applying the formulas* (15) *and* (16).

As a spin-off, this argument provides a proof of the formula (6) for the number of different keyblock orientation patterns—the number of pairs of distinct indices i, j different from a given index k is of course $\binom{m-1}{2}$.

This result leads immediately to the following efficient algorithm for generating the full list of keyblock orientation patterns:

With $k=1$ held fixed, successively reduce the system of equations (11) to the form (14) for each pair of distinct indices i, j between 2 and m inclusive, and apply formulas (15) and (16) to generate orientation patterns. The result will be the complete list of keyblock orientation patterns. The successive reductions of (11) can be performed incrementally using so-called pivot operations in a linear programming tableau.

It remains to show that in (14) none of the coefficients or right-hand sides can vanish. Let

$$\sum_{i=1}^{m} a_i\lambda_i = b$$

represent any nontrivial equation deduced by reduction operations from (11). This means that this equation is some linear combination of the three component equations in (11), which implies the existence of a nonzero vector $\mathbf{a}$ such that

$$a_i = \mathbf{a} \cdot \mathbf{n}_i, \qquad i = 1, \cdots, m \text{ and } b = \mathbf{a} \cdot \mathbf{N}.$$

Since no trio of the normals is coplanar, clearly no more than two of the coefficients a_i, b can vanish. In (14) this leads to the desired conclusion.

6. Keyblocks visible in more than one excavation face of a convex excavation. The above considerations do not apply to blocks of rock that, instead of being visible only in some single excavation face, are visible in two or more excavation faces. We limit attention to the case in which the relevant part of the excavation is given by simultaneous inequalities of the form

$$\text{(20)} \qquad \mathbf{N}_a \cdot \mathbf{r} \geqq d_a, \qquad a = 1, 2, \cdots, n,$$

where part of the block is visible in each of these n faces. The block is thus given by conditions of the form

$$\text{(21)} \qquad \begin{aligned} e_i\mathbf{n}_i \cdot \mathbf{r} &\geqq e_ic_i \quad \text{for } \textit{all } i = 1, 2, \cdots, m, \\ \mathbf{N}_a \cdot \mathbf{r} &\leqq d_a \quad \text{for } \textit{some } a = 1, 2, \cdots, n. \end{aligned}$$

As before, the block is movable if and only if there exists a vector $\mathbf{v}$ such that (7) holds. Turning to the question of the boundedness of the block, let us denote by B_a the set of all points $\mathbf{r}$ such that

$$e_i\mathbf{n}_i \cdot \mathbf{r} \geqq e_ic_i \quad \text{for all } i = 1, \cdots, m \text{ and } \mathbf{N}_a \cdot \mathbf{r} \leqq d_a.$$

Then the block itself is the union of the sets B_a and hence is bounded if and only if each of the sets B_a is bounded. Note that each set B_a is nonempty since the block is assumed to be visible in each of the excavation faces. Since (7) is independent of any particular face, we arrive thus at the conclusion that the block (21) is bounded and movable if and only if each of the blocks B_a is bounded and movable. It follows from the multiplier characterisation that *the block is a keyblock in a convex excavation if and only if its orientation pattern is such that for every excavation face in which it is visible, the system* (11) *with sign conditions* (12) *has a nonempty bounded set of solutions.* In applying the algorithm described in the previous paragraph, the different normal vectors $\mathbf{N}_a$ involved may be treated simultaneously in performing the reduction of and subsequent pivot operations on the system (11). An orientation pattern then corresponds to a keyblock if and only if it appears in the list corresponding to each choice of normal vector $\mathbf{N}_a$.

7. Hanging keyblocks. In practice a keyblock may present a particular danger if it is potentially capable of sliding out in a direction having a *downward* component. Such keyblocks will be termed hanging keyblocks, and a block (21) is a hanging keyblock if it is bounded and there exists a translation vector $\mathbf{v}$ such that

$$\begin{aligned} &e_i\mathbf{n}_i\cdot\mathbf{v}>0 \quad \text{for all } i=1,2,\cdots,m,\\ &\mathbf{K}\cdot\mathbf{v}\leqq 0, \end{aligned} \tag{22}$$

where $\mathbf{K}$ denotes a vector pointing vertically upwards. We assume that no two normal vectors $\mathbf{n}_i$ are coplanar with $\mathbf{K}$. Once again, by Tucker's Theorem of the Alternative, this is equivalent to the inconsistency of the system

$$\mu\mathbf{K}=\sum_{i=1}^{m}\mu_ie_i\mathbf{n}_i,\quad \mu\geqq 0,\quad \mu_i>0 \quad \text{but not all zero} \tag{23}$$

and the individual consistency of each of the systems (for $a=1,2,\cdots,n$)

$$\mathbf{N}_a=\sum_{\mu=1}^{m}\mu_ie_i\mathbf{n}_i,\quad \mu_i>0. \tag{24}$$

Suppose this holds. Then, in particular, (23) with $\mu=0$ is inconsistent, which is precisely the condition for the boundedness of the solution sets of (24). Thus the algorithm of § 5 will include the particular orientation pattern $(e_1,\cdots,e_n)$ in the list generated for each choice $\mathbf{N}=\mathbf{N}_a$, but not for the choice $\mathbf{N}=\mathbf{K}$. Conversely, if this outcome occurs, then each of the systems (24) has a nonempty bounded solution set, while that of (23) must be either nonempty but unbounded, or empty. However, the condition for boundedness of solution sets in (23) is the same as that for (24), i.e. (23) with $\mu=0$ to be inconsistent. Hence the first alternative here cannot occur, and (23) must have been inconsistent.

Thus a block (21), *visible in n faces of a convex excavation, is a hanging keyblock if and only if its orientation pattern appears in the list generated by the algorithm of* § 5 *for each of the specifications* $\mathbf{N}=\mathbf{N}_a$, *but not in the list generated for* $\mathbf{N}=\mathbf{K}$.

Once again, in executing the reductions of system (11), the various specifications for $\mathbf{N}$ can be treated simultaneously.

8. Example. For illustrative purposes we use the same example as is treated by Shi' and Goodman in [6]. This involves $m=4$ sets of discontinuity planes and one excavation face, the orientations of which are given in Table 1. In addition, we consider a second excavation face, viz. a horizontal roof with inward normal vector directed vertically downwards.

TABLE 1

	Dip direction in °E of N	Dip angle in ° below horizontal
Family 1	172°	71°
Family 2	243°	68°
Family 3	302°	38°
Family 4	343°	13°
Excavation Face (excavation above this face)	300°	60°

Resolving the various upward normal vectors into components east, north, and vertically upwards, equations (11) become

$$\begin{bmatrix} .132 & -.826 & -.522 & -.066 \\ -.936 & -.421 & .326 & .215 \\ .326 & .375 & .788 & .974 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \begin{bmatrix} -.750 & 0 \\ .433 & 0 \\ .500 & -1 \end{bmatrix},$$

where we propose to deal simultaneously with two selections of excavation face. The first step is to select λ_1 to play the special role in the algorithm, and to reduce this system of equations to the form (14) for the first pair of indices $i, j \neq 1$, say $i=2, j=3$. This yields the equivalent system

$$\begin{bmatrix} 1 & 0 & 0 & .539 \\ 0 & 1 & 0 & -.679 \\ 0 & 0 & 1 & 1.337 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \begin{bmatrix} -.455 & -.758 \\ .451 & .691 \\ .608 & -1.285 \end{bmatrix}. \tag{25}$$

To each of the right-hand sides separately we now apply formulas (15) and (16). Dealing first with the sloping excavation face, we have from (15)

$$e_1 = \operatorname{sgn}(-.455) = -1, \quad e_2 = \operatorname{sgn}(.451) = +1, \quad e_3 = \operatorname{sgn}(.608) = +1,$$

while from (16)

$$e_4 = e_1 \operatorname{sgn}(.539) = -1.$$

Thus the orientation pattern $(-1, +1, +1, -1)$ corresponds to a keyblock. This means that a block that is bounded above by planes of the first and fourth families and below by planes of the second and third, and that is visible only in the sloping excavation face, is a keyblock. For this reason we use the simpler notation $L\ U\ U\ L$ for the orientation pattern.

Similarly, for the second right-hand side in (25) we obtain

$$e_1 = -1, \quad e_2 = +1, \quad e_3 = -1, \quad e_4 = -1, \quad \text{or } L\ U\ L\ L.$$

In tableau form the system (25) is simply

	λ_4		
λ_1	.539	−.455	−.758
λ_2	−.679	.451	.691
λ_3	1.337	.608	−1.285

and by a simple "pivot" operation we bring the variable λ_4 into the basis in exchange for λ_3, in order to deal with the index pair $i=2, j=4$. This produces the new tableau

	λ_3		
λ_1	$-.403$	$-.700$	$-.240$
λ_2	$.508$	$.760$	$.038$
λ_4	$.748$	$.455$	$-.961$

From (15) and (16) the corresponding orientation pattern for the first right-hand side is

$$e_1=\operatorname{sgn}(-.700)=-1, \quad e_2=\operatorname{sgn}(.760)=+1, \quad e_4=\operatorname{sgn}(.455)=+1,$$
$$e_3=e_1\operatorname{sgn}(-.403)=+1, \quad \text{or } L\ U\ U\ U,$$

and similarly, $L\ U\ U\ L$ for the second right-hand side.

A final pivot operation to swap λ_2 and λ_3 produces the tableau

	λ_2		
λ_1	$.793$	$-.097$	$-.210$
λ_3	1.969	1.496	$.069$
λ_4	-1.472	$-.664$	-1.017

From (15) and (16), the corresponding orientation patterns are $L\ L\ U\ L$ for both right-hand sides. The complete list for the sloping excavation face is thus

$$(26) \qquad \begin{array}{cccc} L & U & U & L \\ L & U & U & U \\ L & L & U & L, \end{array}$$

which agrees with that given by Shi and Goodman in [6, Fig. 7a]. For the horizontal roof the complete list is

$$(27) \qquad \begin{array}{cccc} L & U & L & L \\ L & U & U & L \\ L & L & U & L, \end{array}$$

and any block that is visible solely in this roof and has any one of these orientation patterns is a keyblock.

The patterns common to both lists are

$$(28) \qquad \begin{array}{cccc} L & U & U & L \\ L & L & U & L, \end{array}$$

and these correspond, as shown in § 6, to keyblocks that embrace the corner between the sloping excavation face and the roof.

We can determine which of these are hanging keyblocks by observing that replacement of the vertically downward normal vector by the vertically upward one (i.e. by **K**) merely reverses the sign of all orientations for that case. Thus the list corresponding to $\mathbf{N}=\mathbf{K}$ can be deduced from (27) to be

$$\begin{matrix} U & L & U & U \\ U & L & L & U \\ U & U & L & U. \end{matrix} \tag{29}$$

According to § 7, patterns that are in the list (26) but not in (29) correspond to hanging keyblocks—in this case all three do. Likewise, all three patterns in (27), and hence also those in (28), also correspond to hanging keyblocks.

Finally, we observe that if the excavation is below, rather than above, the sloping excavation face, so that (26) is replaced by the list

$$\begin{matrix} U & L & L & U \\ U & L & L & L \\ U & U & L & U, \end{matrix} \tag{30}$$

then there is no overlap with (27), so that in this case no blocks that intersect the corner are keyblocks. The pattern $U\ L\ L\ L$ is the only one from (30) that is not in (29), and hence it represents the only hanging keyblock in this case.

Appendix—Statement of Tucker's Theorem of the Alternative. For the convenience of readers we give a statement of Tucker's Theorem of the Alternative, as presented in [3, p. 29].

Let B, C and D be given matrices, with B being nonvacuous. Then either the system $Bx \geqq 0$ but not all zero, $Cx \geqq 0$, $Dx = 0$ has a solution x or the system $\lambda' B + \mu' C + \nu' D = 0$, $\lambda > 0$, $\mu \geqq 0$ has a solution λ, μ, ν, but never both.

This theorem was first given by Tucker in [7].

Acknowledgments. The authors are grateful to Messrs W. N. Merten, W. L. van Heerden and F. A. Vreede of the National Mechanical Engineering Research Institute of the CSIR, Pretoria for stimulating discussions on the practical importance of the Shi-Goodman work, and for prompting our interest in in the keyblock identification problem.

REFERENCES

[1] S. A. Cartney, *The ubiquitous joint method*, in Cavern Design at Dinorwic Power Station, Tunnels and Tunneling, 1977, pp. 54-57.

[2] G. F. Hadley, *Linear Programming*, Addison-Wesley, Reading, MA, 1962.

[3] O. L. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[4] R. T. Rockafellar, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.

[5] G. H. Shi and R. E. Goodman, *Underground support design using block theory to determine keyblock bolting requirements*, Proc. SANGORM Symposium on Rock Mechanics in the Design of Tunnels, Pretoria, 1963, pp. 81-105.

[6] ———, *A new concept for support of underground and surface excavations in discontinuous rocks based on a keystone principle*, Proc. 22nd U.S. Symposium on Rock Mechanics, Massachusetts Institute of Technology, Cambridge, MA, 1981, pp. 290-296.

[7] A. W. Tucker, *Dual systems of homogeneous linear relations*, in H. W. Kuhn and A. W. Tucker, Linear Inequalities and Related Systems, Annals of Mathematics Studies No. 38, Princeton Univ. Press, Princeton, NJ, 1956.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

017

HARD ENUMERATION PROBLEMS IN GEOMETRY AND COMBINATORICS*

NATHAN LINIAL†

Abstract. A number of natural enumeration problems in geometry and combinatorics are shown to be complete in the class #P introduced by Valiant. Among others this is established for the numeration of vertices and of facets of a polytope, acyclic orientations of a graph and satisfying assignments of implicative boolean formulas.

Key words. #P, polytopes, partial orders, acylic orientations

AMS(MOS) subject classifications. 05C20, 52A25, 68C25

Introduction. This article contains a contribution to the theory of hard enumeration problems. The foundations of this area were laid by Valiant [Va1], [Va2] who defined the class #P of enumeration problems and the subclass of problems complete in #P. The most interesting of his results is the #P completeness of computing permanents of 0–1 matrices. This problem can also be stated as the problem of enumerating perfect matchings in bipartite graphs. While deciding whether a bipartite graph has a perfect matching can be done in polynomial time [H] the enumeration problem is #P-complete.

Valiant's pioneering work was continued by a recent article of Provan and Ball [PB] who prove the #P-completeness of a number of natural enumeration problems. With every enumeration problem there is an associated decision problem. Instead of asking for the number of objects in question we ask whether this number is zero or not. The decision problem associated with the computation of the permanent function is the question whether a given bipartite graph has a perfect matching. While this decision is solvable in polynomial time, it is by no means trivial. Notice, however, that for many of the problems discussed in [PB] the situation is even more extreme: Consider for example the problem of enumerating independent sets in a bipartite graph, the decision problem associated with this enumeration problem is trivial: Every graph has an independent set of vertices. So an enumeration problem can be #P-complete even if the existence problem is trivial.

This article assumes acquaintance with the theory of #P-completeness as presented in [Va1], [PB] and [GJ]. Our purpose is to present a number of natural enumeration problems which belong to the class of #P-complete problems. The problems are geometric, combinatorial and from propositional calculus.

Here is our main theorem:

THEOREM. *The following enumeration problems are* #P-*complete.*

(1) *Vertices in a polytope.*
Input: A system of linear inequalities $Ax \leqq b$ defining a polytope $P \subseteq \mathbb{R}^n$.
Output: The number of vertices of P.

(2) *d-dimensional faces of a polytope (fixed d).*
Input: As in (1).
Output: The number of d-dimensional faces of P.

* Received by the editors December 10, 1984.

† Institute of Mathematics and Computer Science, Hebrew University, Jerusalem 91904, Israel.

(3) *Facets of a polytope.*
Input: A finite set of points in $\mathbb{R}^n$.
Output: The number of facets ($(n-1)$-dimensional faces) of P.

(4) *Components of slotted space.*
Input: A set $\{H_i | i \in I\}$ of hyperplanes in $\mathbb{R}^n$.
Output: The number of connected components of $\{\mathbb{R}^n \backslash \bigcup H_i | i \in I\}$.

(5) *Acyclic orientations of a graph.*
Input: A graph $G = (V, E)$.
Output: The number of orientations of G with no directed circuit.

(6) 3*-colorings of a bipartite graph.*
Input: A bipartite graph $G = (A, B, E)$.
Output: The number of ways to properly color G with 3 colors.

(7) *Satisfying assignments of an implicative Boolean formula.*
Input: A Boolean formula B on variables $x_1, \cdots, x_n$ of the form $B = \bigwedge_{i=1}^{t} (x_{i1} \vee \bar{x}_{i2})$.
Output: Number of truth assignments for $x_1, \cdots, x_n$ which makes B true.

Proof. (1) We use the fact from [PB] that enumerating order ideals is #P-complete. Given a poset $(P, \geqq)$ with $P = [n]$ we associate with P a polytope $B = B(P)$ in $\mathbb{R}^n$ as follows:

$$B = \{x \in \mathbb{R}^n | 1 \geqq x_i \geqq 0, x_i > x_j \text{ if } i \geqq j \text{ in } P\}.$$

(See [St2], [Li], [KS] where use is made of this polytope.)

We claim that the vertices of B are in 1:1 correspondence with the order ideals of $(P, \geqq)$. First we prove that all vertices of B have 0–1 coordinates. Let $x \in B$ have some $0 < x_i < 1$. If $\alpha = \max\{x_j | 0 < x_j < 1\}$, then by replacing all coordinates $x_j = \alpha$ by $\alpha + \varepsilon$ or by $\alpha - \varepsilon$ we will get a point of B. This implies that x is not a vertex of B. The correspondence between vertices and ideals is as follows:

$$x \in \text{vert}(B) \leftrightarrow S = \{1 \leqq j \leqq n | x_j = 0\}.$$

It is easily verified that S is an ideal and that this correspondence is bijective.

(2) Suppose that for some fixed d we can find $f_d(K)$ the number of d-dimensional faces of a polytope K. Consider r-fold pyramids P_r with K as basis. In [Gru, p. 55] one finds

$$(*) \qquad f_d(P_r) = \sum_i \binom{r}{i} f_{d-i}(K).$$

If we write (*) for $r = 0, \cdots, d$ and have all $f_d(P_r)$ evaluated, then we obtain a system of equations in unknown

$$f_0(K), \cdots, f_k(K).$$

This system of equations has a triangular matrix and so they can be solved successively and $f_0(K)$ can be determined in polynomial time. Since evaluating $f_0(K)$ = the number of vertices of K is #P-complete by (1), our claim follows.

(3) This is just the dual of (1): See [Gru, p. 46] for polytope duality.

(5) The proof here is based on two observations.

PROPOSITION [St1]. *Let $G = (V, E)$ be a graph with n vertices and let $P(G, \lambda)$ be its chromatic polynomial. Then $(-1)^n P(G, -1)$ equals the number of acyclic orientations of G.*

For the other observation we have to define the operation of *join* of two graphs $G=(V_1, E_1)$, $H=(V_2, E_2)$ where $V_1 \cap V_2 = \emptyset$. The join $G+H$ has $V_1 \cup V_2$ as its vertex set and

$$E_1 \cup E_2 \cup \{[x, y] | x \in V_1, y \in V_2\}$$

as its edge set. The following observation is immediate.

PROPOSITION. $P(G+K_t, \lambda) = \lambda(\lambda-1), \cdots, (\lambda-t+1)P(G, \lambda-t)$.

Now we can combine these two facts as follows. Being able to enumerate acyclic orientations is equivalent to computing $P(G, -1)$ for the graph. But if we have the values of $P(G+K_t, -1)$ for $t=1, \cdots, n$, that means we can calculate the integers

$$P(G, -j) \qquad (n+1 \geqq j \geqq 2).$$

But P is a monic polynomial of degree n so from these numbers we can compute $P(G, \lambda)$, the chromatic polynomial of G. This is a #P-complete problem because the reduction to coloring is parsimonious [GJ, p. 169].

(4) The proof here makes use of (5) that enumerating acyclic orientations is #P-complete and on the following result of Greene. $H_{ij} \subseteq \mathbb{R}^n$ is the hyperplane given by $\{x \in \mathbb{R}^n | x_i = x_j\}$.

PROPOSITION [Gre]. *Let $G=(V, E)$ be a graph on n vertices and consider*

$$S(G) = \mathbb{R}^n \backslash \cup H_{ij}$$

where the union is over all i, j such that $[i, j] \in E$. The number of connected components of $S(G)$ equals the number of acyclic orientations of G.

(6) We base this proof on the #P-completeness of enumerating independent sets in bipartite graphs [PB]. Let $G=(A, B, E)$ be a partite graph for which we want to find $I(G)$ the number of independent sets. Consider a graph H which is obtained by adding two new vertices a, b with a being adjacent to all vertices in $B \cup \{b\}$ and b to all vertices of $A \cup \{a\}$. Now let us compute $\chi(H, 3)$, the number of 3-colorings of H. Suppose w.l.o.g. that a, b are colored 1, 2 respectively. The 3-coloring is now uniquely defined by the set of vertices colored 3. This can be any independent set of G and so

$$\chi(H, 3) = 6I(G).$$

This proves the #P-completeness of computing $\chi(H, 3)$.

(7) This follows from #P-completeness of enumerating ideals in posets [PB]: Let $(P, \geqq)$ be a poset with $P=\{p_1, \cdots, p_n\}$. Associate with it the Boolean expression

$$B = \wedge\{x_i \vee \bar{x}_j | p_j > p_i \text{ in } P\}.$$

It is fairly easy to verify that the set of x_i which are assigned a true value in any assignment satisfying B is an ideal in $(P, \geqq)$ and that all ideals are obtained in this way.

Let us mention in closing a most intriguing problem in this field: For a poset $(P, \geqq)$ a *linear extension* is a 1:1 mapping $f: P \to \{1, \cdots, |P|\}$ such that if $x<y$ in P then $f(x)<f(y)$. Consider the problem:

Enumeration of linear extensions.
Input: A poset $(P, \geqq)$.
Output: $L(P)$, the number of linear extensions of $(P, \leqq)$.

Conjecture. The enumeration of linear extensions is a #P-complete problem.

A proof of this conjecture will provide a first explicit statement to the effect that computing the volume of a convex polytope is a hard computational problem. To see this we remind the reader about the polytope $B(P)$ which was used in proving part 1 of our main theorem. We quote without proof of the following fact from [Li]:

PROPOSITION. *For a poset $(P, \geqq)$ on $|P| = p$ elements $L(P)$ the number of linear extensions of P satisfies*

$$L(P) = p!\,\mathrm{vol}\,(B(P)).$$

The connection between the #P-completeness of enumerating linear extensions and the complexity of evaluating the volume of a convex polytope is now clear.

Let us also comment about the relationship between the number of linear extensions of a poset and enumerating order ideals. We use $I(P)$ to denote the number of ideals in the poset P. For posets P, Q we define their *product* $P \times Q$ to be a partial order on the cartesian product of P and Q with $(x_1, y_1) \geqq (x_2, y_2)$ if $x_1 \geqq x_2$ in P and $y_1 \geqq y_2$ in Q. A mapping $f: P \to Q$ is *order preserving* if $x \geqq y$ in P implies $f(x) \geqq f(y)$ in Q.

PROPOSITION. *Let $(P, \geqq)$ be a poset and let C_t be the chain on t elements. Then $I(P \times C_t)$ equals the number of order preserving maps $f: P \to \{0, 1, \cdots, t\}$.*

Proof. With an ideal $J \subseteq P$ we associate a function $f: P \to \{0, \cdots, t\}$ as follows: For every $x \in P$ there is unique $t \geqq j \geqq 0$ such that $(x, j) \in J$ and $(x, j-1) \notin J$. Let $f(x) = t - j$ for that value j. Since J is an ideal, f is well defined and easily seen to be order preserving. It is also a routine matter verifying that this correspondence is bijective.

Now we come to the expression for the number of linear extensions of a poset.

THEOREM. *For a poset $(P, \geqq)$ on $|P| = n$ elements, the number of linear extensions $L(P)$ satisfies*

$$L(P) = I(P \times C_{n-1}) - nI(P \times C_{n-2}) + \binom{n}{2} I(P \times C_{n-3})$$
$$- + \cdots \pm \binom{n}{n-2} I(P \times C_1) \mp \binom{n}{n-1}.$$

Proof. This follows from the previous proposition and Inclusion-Exclusion. Classify order preserving maps $f: P \to \{0, \cdots, n-1\}$ according to their range. There are $I(P \times C_{n-1})$ such mappings altogether. Say f has property t $(n-1 \geqq t \geqq 0)$ if t is not in the range of f. $L(P)$ is the number of order preserving maps which are onto, i.e., have no property and there are

$$\binom{n}{j} I(P + C_{n-j-1})$$

maps having a given set of j properties.

REFERENCES

[Gre] C. GREENE, *Acyclic orientations (Notes)*, in Higher Combinatorics, M. Aigner, ed., D. Reidel, Dordrecht, 1977, pp. 65-68.

[Gru] B. GRÜNBAUM, *Convex Polytopes*, Wiley Interscience, New York, 1967.

[GJ] M. GAREY AND D. JOHNSON, *Computers and Intractability: A Guide to the Theory of* NP-*Completeness*, W. H. Freeman, San Francisco, 1979.

[H] M. HALL, JR., *An algorithm for distinctive representatives*, Amer. Math. Monthly, 63 (1956), pp. 716-717.

[KS] J. KAHN AND M. E. SAKS, *Every poset has a good comparison*, Combinatorica, to appear.

[PB] S. PROVAN AND M. O. BALL, *On the complexity of counting cuts and of computing the probability that a graph is connected*, SIAM J. Comput., 12 (1983), pp. 777-788.

[Li] N. LINIAL, *The information theoretic bound is good for merging*, SIAM J. Comput., to appear.

[St1] R. P. STANLEY, *Acyclic orientations of graphs*, Discrete Math., 5 (1973), pp. 171–178.

[St2] ———, *Two combinatorial applications of the Alexandrov Fenchel inequalities*, J. Combin. Theory, ser. A, 31 (1981), pp. 56–65.

[Va1] L. G. VALIANT, *The complexity of computing the permanent*, Theor. Comput. Sci., 8 (1979), pp. 189–201.

[Va2] ———, *The complexity of enumeration and reliability problems*, SIAM J. Comput., 8 (1979), pp. 410–421.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 2, April 1986

018

ERRATUM: VOLTERRA MULTIPLIERS II*

RAY REDHEFFER†

The following note should be added on page 621.

A more accurate calculation by Professor Kachar of Karlsruhe University indicates that the inequality $c > 12.58$ following (28) should be $c \geqq 12.60043$, leading to minor emendation elsewhere. Meanwhile the method of this paper has been programmed on a computer by Wolfgang Walter, Jr., also of Karlsruhe University. He finds that the multiplier exists if $c \geqq 12.45$ and not if $c \leqq 12.40$.

* This Journal, 6 (1985), pp. 612-623.

† Department of Mathematics, University of California, Los Angeles, California 90024.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

001

A PARALLEL BLOCK ITERATIVE SCHEME APPLIED TO COMPUTATIONS IN STRUCTURAL ANALYSIS*

ROBERT J. PLEMMONS†

Abstract. In this paper it is shown how a block cyclic successive overrelaxation direct-iterative method can be applied to the parallel solution of certain large-scale linear equality-constrained quadratic programming problems. The scheme is similar in nature to those studied recently by de Pillis, Niethammer and Varga and by Markham, Neumann and Plemmons for solving large sparse least squares problems. It is based upon a partitioning strategy of the fundamental matrix into a block consistently ordered 2-cyclic form where the nonzero eigenvalues of the Jacobi matrix are all pure imaginary. The method is shown to be globally convergent and convergence rates are established.

Applications of the algorithm are discussed for large-scale structural analysis computations where it is shown how the algorithm can be adapted to the simultaneous computation of the system forces and the nodal displacements. Here, advantage can be taken of the special forms of the matrices involved. In particular, it is shown that much of the algorithm lends itself to efficient implementation on pipelined vector machines and on multiprocessors.

Key words. block successive overrelaxation, constrained minimization, linear systems, parallel processing, structural analysis

AMS(MOS) subject classifications. Primary: 49D40, 65F10, 65N20

1. Introduction. The first part of this paper is concerned with a class of iterative schemes for solving the following constrained minimization problem:

$$\begin{aligned} &\text{Minimize } \tfrac{1}{2}x^TAx - x^Ts \\ &\text{subject to } Ex = t. \end{aligned} \tag{1.1}$$

Here it is assumed that A is a real positive semidefinite $n \times n$ matrix, E is a real $m \times n$ matrix with full row rank m, t is a real m-vector and s is a real n-vector. Thus (1.1) is an equality-constrained quadratic programming problem. It is known from the theory of quadratic programming (e.g., Gill and Murray [1974] or Hadley [1964]) that if A and E have no nontrivial null vectors in common then (1.1) has a unique solution x which forms part of the solution $\binom{x}{\lambda}$ to the following system of linear equations.

$$\begin{bmatrix} A & E^T \\ E & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} s \\ t \end{bmatrix}. \tag{1.2}$$

Under the assumptions made above, the coefficient matrix

$$B = \begin{bmatrix} A & E^T \\ E & 0 \end{bmatrix} \tag{1.3}$$

of (1.2) is nonsingular. The system (1.2) is sometimes called the *fundamental system*, B is called the *fundamental matrix* and λ the vector of *Lagrange multipliers* for the quadratic programming problem.

Iterative schemes for solving (1.2) have recently drawn considerable interest. Dyn and Ferguson [1983] have considered methods based upon classical splittings of the matrix A. Axelsson [1984] has considered various implementations of the conjugate

* Received by the editors March 18, 1985, and in revised form June 17, 1985.

† Departments of Computer Science and Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205. This research was sponsored by the U.S. Air Force under grant AFOSR-83-0255 and by the National Science Foundation under grant DMS-82-19500-A01.

gradient method applied to (1.2). These types of schemes are typically efficient only when the matrix A is large and sparse and there are only a moderate number of constraints.

In the first part of this paper we develop a direct-iterative scheme for solving (1.2) which is based upon a partitioning strategy of the fundamental matrix into a block consistently ordered 2-cyclic form, where the nonzero eigenvalues of the Jacobi matrix are all pure imaginary. In contrast to the scheme of Dyn and Ferguson, both the matrices E and A are involved in the splitting. This work was motivated in part by the work of Markham, Neumann and Plemmons [1985] in which a similar 2-cyclic scheme was used to solve certain large sparse linear least squares problems.

It is assumed here that A and E can be permuted or partitioned so that A is block diagonal and that E has the form $[E_1\ \ E_2]$, with E_1 square and nonsingular. That these assumptions hold for a large class of engineering problems is established in § 3. Here an implementation of the algorithm is given for solving large-scale structural analysis problems on pipelined vector machines and on multiprocessors. The 2-block SOR iterative method is developed and its convergence is established in the next section.

2. Block SOR iteration. Consider the constrained minimization problem (1.1) with A $n\times n$ and positive semidefinite and E $m\times n$ with rank m. Assume first that A and E have been permuted or partitioned so that A has the block diagonal form

$$A=\begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$$

with A_1 $m\times m$ and with A_2 nonsingular and so that $E=[E_1\ \ E_2]$ with E_1 $m\times m$ and nonsingular. Partition x and s conformally into $x=[\begin{smallmatrix}x_1\\x_2\end{smallmatrix}]$ and $s=[\begin{smallmatrix}s_1\\s_2\end{smallmatrix}]$ with x_1 and s_1 m-vectors. Then the $(m+n)\times(m+n)$ fundamental system of linear equations (1.2) can be expressed as

$$\begin{bmatrix} A_1 & 0 & E_1^T \\ 0 & A_2 & E_2^T \\ E_1 & E_2 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \lambda \end{bmatrix}=\begin{bmatrix} s_1 \\ s_2 \\ t \end{bmatrix}.$$

By interchanging row blocks one and three the following block system is obtained

$$\begin{bmatrix} E_1 & E_2 & 0 \\ 0 & A_2 & E_2^T \\ A_1 & 0 & E_1^T \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \lambda \end{bmatrix}=\begin{bmatrix} t \\ s_2 \\ s_1 \end{bmatrix}. \tag{2.1}$$

Then

$$C\equiv\begin{bmatrix} E_1 & E_2 & 0 \\ 0 & A_2 & E_2^T \\ A_1 & 0 & E_1^T \end{bmatrix} \tag{2.2}$$

is block 3-cyclic (see Varga [1962, pp. 99–114]).

At this point a natural 3-block SOR iterative scheme could be developed for solving (2.1), perhaps in a manner similar to the 3-block scheme studied in Chen [1975], Plemmons [1979] and de Pillis, Niethammer and Varga [1984] for solving large sparse linear least squares problems. It was pointed out by Markham, Neumann and Plemmons [1985], however, that a repartitioning of the coefficient matrix into a 2-block form leads to a method which always converges and has superior convergence rates. Although the matrix C in (2.2) is somewhat more complicated than the coefficient matrix for the least squares problem, an analysis similar to that given in Markham, Neumann and Plemmons [1985] can be given to show that a 2-block partitioning of C also leads to superior convergence rates for solving the constrained minimization problem (1.1). (The least squares problem can, in fact, be considered as a special case of the problem discussed here.) Thus only the 2-block approach will be developed in this paper.

The coefficient matrix C in (2.2) can be partitioned into a 2×2 block form in two obvious ways, each leading to similar convergence results. The following partitioning is chosen.

$$C=\left[\begin{array}{cc|c} E_1 & E_2 & 0 \\ 0 & A_2 & E_2^T \\ \hline A_1 & 0 & E_1^T \end{array}\right]. \tag{2.3}$$

This leads to the following 2-block SOR direct-iterative scheme applied to the solution to (1.1). Let D, L and U denote the 2-block matrices

$$D=\left[\begin{array}{cc|c} E_1 & E_2 & 0 \\ 0 & A_2 & 0 \\ \hline 0 & 0 & E_1^T \end{array}\right],\quad L=\left[\begin{array}{cc|c} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \hline -A_1 & 0 & 0 \end{array}\right],\quad U=\left[\begin{array}{cc|c} 0 & 0 & 0 \\ 0 & 0 & -E_2^T \\ \hline 0 & 0 & 0 \end{array}\right]. \tag{2.4}$$

Let

$$y\equiv\begin{bmatrix} x_1 \\ x_2 \\ \lambda \end{bmatrix},\qquad b\equiv\begin{bmatrix} t \\ s_2 \\ s_1 \end{bmatrix}. \tag{2.5}$$

Then the 2-block SOR iterative scheme for solving the constrained minimization problem (1.1) can be expressed in the following matrix form:

$$y^{(k+1)}=\mathscr{L}_\omega y^{(k)}+(D-\omega L)^{-1}\omega b,\qquad k=0,1,\cdots \tag{2.6}$$

where ω is the SOR parameter and the SOR iteration matrix is given by

$$\mathscr{L}_\omega=(D-\omega L)^{-1}[(1-\omega)D+\omega U]. \tag{2.7}$$

A detailed block by block version of the iteration (2.6) will be provided after the following convergence theorem is established. Here ρ denotes the spectral radius.

THEOREM 1. *For the constrained minimization problem* (1.1) *let A and E be such that A_2 is positive definite and E_1 is nonsingular in the block partitioned form* (2.3). *Let*

$$\rho\equiv\rho[(E_1^{-1}E_2)A_2^{-1}(E_1^{-1}E_2)^TA_1]. \tag{2.8}$$

Then the 2-block SOR method (2.6) *converges to the solution $y=[{x \atop \lambda}]$ to the fundamental system* (1.2) *for every $y^{(0)}$ for all relaxation parameters ω in the interval*

$$0<\omega<\frac{2}{1+\rho}. \tag{2.9}$$

Here the optimum SOR relaxation parameter $\omega_b=\omega(\rho)$ is given by

$$\omega_b=\frac{2}{\sqrt{1+\rho^2}+1} \tag{2.10}$$

and the spectral radius of the resulting iteration matrix $\mathcal{L}_{\omega_b}$ *is given by*

$$\rho(\mathcal{L}_{\omega_b}) = \left[\frac{\rho}{\sqrt{1+\rho^2}+1}\right]^2, \tag{2.11}$$

so that the method converges for all ρ *given by* (2.8).

Proof. An adaptation of the theory of consistently ordered p-cyclic matrices will be applied. The proof will be obtained from a simple application of the results in Young [1971, § 6.4].

Observe first that the Jacobi iteration matrix J for the 2-block form of C in (2.3) is given by

$$\begin{aligned} J = D^{-1}(L+U) &= \begin{bmatrix} \begin{pmatrix} E_1 & E_2 \\ 0 & A_2 \end{pmatrix}^{-1} & \begin{matrix} 0 \\ 0 \end{matrix} \\ \begin{matrix} 0 & \quad 0 \end{matrix} & E_1^{-T} \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -E_2^T \\ -A_1 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & E_1^{-1}E_2A_2^{-1}E_2^T \\ 0 & 0 & -A_2^{-1}E_2^T \\ -E_1^{-T}A_1 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Then

$$J^2 = \begin{bmatrix} -E_1^{-1}E_2A_2^{-1}E_2^TE_1^{-T}A_1 & 0 & 0 \\ A_2^{-1}E_2^TE_1^{-T}A_1 & 0 & 0 \\ & 0 & -E_1^{-T}A_1E_1^{-1}E_2A_2^{-1}E_2^T \end{bmatrix}.$$

Here the first and third diagonal blocks of J^2 have the same eigenvalues. Thus the spectrum of J^2 includes zero and the spectrum of

$$K \equiv -(E_1^{-1}E_2)A_2^{-1}(E_1^{-1}E_2)^TA_1.$$

Since $-K$ is the product of positive semidefinite matrices, all the eigenvalues of J^2 are real and nonpositive. Then J has zero and all pure imaginary eigenvalues and, moreover, $\rho(J) = \rho(K) = \rho$ where ρ is given by (2.8).

Now since C given in (2.3) is block consistently ordered and 2-cyclic, (2.9) follows from Young's Theorem 4.1, p. 191. Finally, since the eigenvalues of J are zero and pure imaginary, (2.10) and (2.11) follow from Young's equations (4.17) on p. 195. □

Observe that with ω given by (2.9) the SOR method always converges, regardless of the value of $\rho(J)$. However, the convergence rate given by (2.11) improves as ρ decreases.

The 2-block SOR scheme given in (2.6) is now summarized.

ALGORITHM 1. *A direct-iterative block cyclic SOR scheme for the constrained minimization problem* (2.1).

Step 1. Factor A_2 and E_1 using appropriate sparse matrix decomposition routines.

Step 2. Estimate $\rho = \rho(J)$ given by (2.8) and the optimum SOR parameter ω_b given by (2.10), using the factorizations in Step 1.

Step 3. Choose initial approximations

$$x^{(0)} = \begin{pmatrix} x_1^{(0)} \\ x_2^{(0)} \end{pmatrix}$$

to the solution x to (1.1) and $\lambda^{(0)}$ to the vector λ of Lagrange multipliers.

Step 4. FOR $k=0$ STEP 1 UNTIL convergence DO

1) Solve

$$A_2x_2^{(k+1)}=(1-\omega_b)A_2x_2^{(k)}-\omega_bE_2^T\lambda^{(k)}+\omega_bs_2. \tag{2.12}$$

2) Solve

$$E_1x_1^{(k+1)}=(1-\omega_b)E_1x_1^{(k)}+E_2[(1-\omega_b)x_2^{(k)}-x_2^{(k+1)}]+\omega_bt.$$

3) Solve

$$E_1^T\lambda^{(k+1)}=(1-\omega_b)E_1^T\lambda^{(k)}-\omega_bA_1x_1^{(k+1)}+\omega_bs_1.$$

The algorithm just described is, of course, a direct-iterative scheme in that it requires the solution of systems of linear equations involving the $m\times m$ and $(n-m)\times(n-m)$ matrices E_1 and A_2, respectively, at each step. However, as indicated earlier, the case of interest here is where E and, accordingly, E_1, have a special structure, such as bandedness. Here well-developed software exists (e.g., Duff and Reid [1979]) for factoring the general unsymmetric matrix E_1 for use in parts 2 and 3 of Step 4. With regard to part 1 of Step 4, eq. (2.12), the $(n-m)\times(n-m)$ matrix A_2 will at least be block diagonal in our applications with relatively small diagonal blocks. For this important case, the equations (2.12) of the iteration can be solved in parallel. Indeed, if A_2 has p diagonal blocks, then (2.12) can be solved in one major time step using p processors in each major iteration. This approach will be developed further in § 3.

Step 2 of the algorithm involves the usual difficult task of effectively estimating the optimum SOR iteration parameter ω_b (which is less than or equal to one so that the process is underrelaxation) and a similar difficulty exists in applying the iterative schemes suggested by Dyn and Ferguson [1983] for solving constrained minimization problems. Some methods for estimating ω_b that apply in our situation are reported in Huang [1983].

Also, Chebyshev acceleration of the Gauss-Seidel implementation of the algorithm ($\omega=1$) might be considered, as for the 3-block formulation discussed in Chen [1975] and Plemmons [1979]. In general, however, the algorithm will not converge for $\omega>2/(1+\rho)$, so that Gauss-Seidel will not converge with $\rho>1$, unless such an acceleration scheme is applied.

It should be remarked that in the direct-iterative method suggested here the formula for ρ in (2.8) simplifies somewhat for certain problems. Such problems are discussed in § 3 where applications of Algorithm 1 to large-scale structural analysis computations are described.

3. Application to structural analysis. One application area of wide interest for the method discussed in this paper involves the engineering analysis of large-scale structures. The *fundamental problem of linear elastic analysis* is that of finding the vector f of internal forces and the vector r of nodal displacements, given a finite element model of a structure and a set of external loads. Specifically, let E be the $m\times n$ equilibrium matrix, p be the m-dimensional vector of nodal (applied) loads and F be the $n\times n$ element-level, block-diagonal element flexibility matrix (here F^{-1} is the element-level force-deformation matrix). Assume that the structure does not form a mechanism (so that rank $E=m$) and the structure is statistically indeterminant with degree of indeterminancy $n-m$.

It is well known (e.g. Robinson [1973]) that the internal force vector f is the particular solution x to the underdetermined system of equilibrium equations

$$Ex = p \tag{3.1}$$

such that

$$\tfrac{1}{2}x^TFx \tag{3.2}$$

is minimal, i.e., f solves (3.1) and satisfies the *principle of minimal energy.* Thus the fundamental problem of linear analysis is the particular constrained minimization problem:

$$\begin{aligned}&\text{Minimize } \tfrac{1}{2}x^TFx\\ &\text{subject to } Ex = p.\end{aligned} \tag{3.3}$$

Clearly then, the methods developed in this paper apply to this problem. In particular, the internal force vector f and the nodal displacement vector r satisfy:

$$\begin{bmatrix} F & E^T \\ E & 0 \end{bmatrix}\begin{bmatrix} f \\ -r \end{bmatrix} = \begin{bmatrix} 0 \\ p \end{bmatrix}, \tag{3.4}$$

so that $-r$ is, in fact, the vector λ of Lagrange multipliers in (1.2).

The most widely used scheme for computing r from (3.4) is the *displacement method.* If block elimination is applied to (3.4), the following $m \times m$ system (e.g., Robinson [1973])

$$Kr = p \tag{3.5}$$

is obtained, where

$$K = EF^{-1}E^T \tag{3.6}$$

and K is called the *stiffness matrix.* One must still recover f from, say,

$$Ff = E^Tr.$$

Here (3.5) might be solved by classical schemes such as sparse Cholesky decomposition or by a pre-conditioned conjugate gradient scheme. In any case the explicit formation of K in (3.6) can worsen the condition of the problem and can lead to a lóss of accuracy, as is the situation for using the normal equations to solve least squares problems.

Another approach is the *force method* where the internal force vector f is computed first. Here, any particular solution y to the equilibrium equations (3.1) is computed and a basis matrix Z for the null space of E is found. Then

$$f = y + Zx$$

where x solves

$$Z^TFZx = -Z^TFy.$$

Thus the force method involves solving the fundamental system

$$\begin{bmatrix} F^{-1} & Z \\ Z^T & 0 \end{bmatrix}\begin{bmatrix} v \\ x \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix}, \tag{3.7}$$

rather than (3.4), where v is the system displacement vector and $f = F^{-1}v$. The matrix Z is called the *self-stress matrix* for the problem (e.g., Robinson [1973]). Although the force method can require considerably more computation for one step than the displacement method, it is useful in reliability analysis and in solving multiple redesign problems when $n-m<m$, where F alone changes in a sequence of problems and Z needs only to be computed once. Some recent sparse matrix schemes for applying the force method are discussed in Kaneko, Lawo and Thierauf [1982] and in Heath, Plemmons and Ward [1984].

Of special interest in this paper is the fact that the element flexibility matrix F is block diagonal in (3.4), i.e.,

$$F=\operatorname{diag}(F_{11},\cdots,F_{qq})$$

where the diagonal blocks F_{ii} are symmetric positive definite and correspond to the ith element in the finite element model of the structure. Generally, these blocks range in size from 1×1 to 6×6 (e.g., Robinson [1973]). This situation facilitates and simplifies considerably the use of F in the 2-block SOR algorithm. Here F is readily partitioned into

$$F=\begin{bmatrix} F_1 & 0 \\ 0 & F_2 \end{bmatrix}$$

where F_1 and F_2 are block diagonal matrices with diagonal blocks having relatively small sizes. Of course for the force method formulation of the problem given in (3.7), F^{-1} is also block diagonal. Here F^{-1} is called the element stiffness matrix and is often available without computing the inverse of F (e.g., Robinson [1973]).

In addition, the equilibrium matrix E for the structure is generally sparse and the ordering of the nodes and elements is often chosen so that E is banded as illustrated in Figs. 1-4. An efficient scheme is described in Berry, Heath, Kaneko, Lawo, Plemmons and Ward [1985] for computing a banded matrix Z for use with (3.7). Moreover, a further ordering of the nodes of the model can lead to a natural partitioning so that $E=[E_1 \;\; E_2]$ where E_1 is a nonsingular banded matrix. Alternatively, a one-pass sparse matrix algorithm for finding column dependencies in E, such as the one given in Heath, Plemmons and Ward [1984], could be used to choose a set of m linearly independent columns to form E_1.

Regarding the implementation of Algorithm 1 for solving structural analysis problems represented by the block equations (3.4) (or (3.7)), advantage can readily

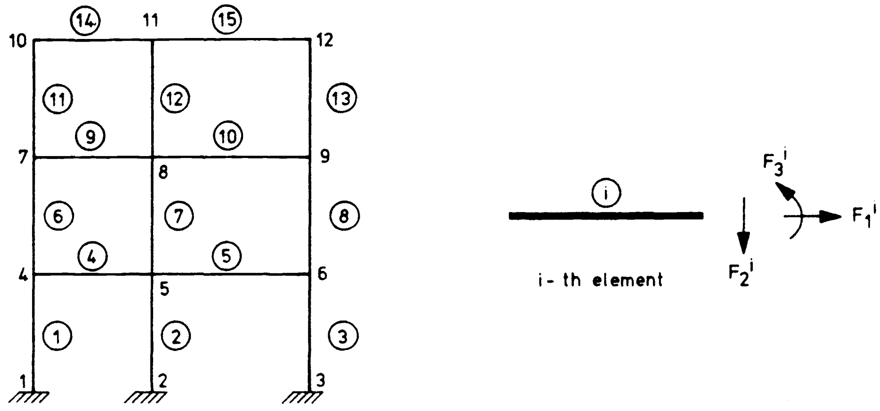


FIG. 1. *Finite element model for two-dimensional frame.*

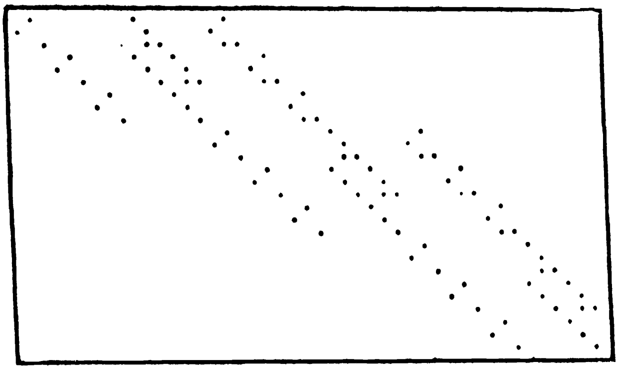


FIG. 2. *Sparsity pattern of equilibrium matrix E for Fig.* 1.

be taken of the special forms of the matrices involved. As indicated above, the matrix E_1 for (3.4) (and the corresponding Z_1 for (3.7)) can generally be taken to be banded, greatly simplifying the implementations of parts 2 and 3 of the iterations and the subsequent solution steps. Furthermore, since F in (3.4) (and the corresponding F^{-1} in (3.7)) is block diagonal, with relatively small diagonal blocks, the equations (2.12) of Part 1 of the iterations lend themselves to parallel implementation.

The application of Algorithm 1 to structural analysis computations is now described. The algorithm is stated for the solution of (3.4) in terms of the system force vector f and the nodal displacement vector r. This corresponds to the displacement method. A similar algorithm can be stated for the solution of (3.7), corresponding to the force (or flexibility) method in structural analysis. It will be assumed, as discussed

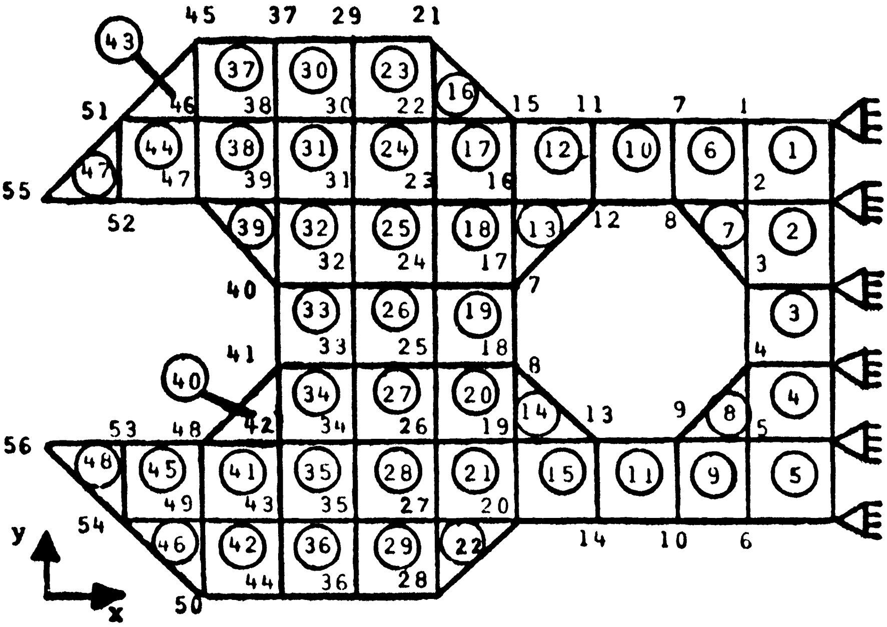


FIG. 3. *Finite element model for plane stress problem* (*wrench*).

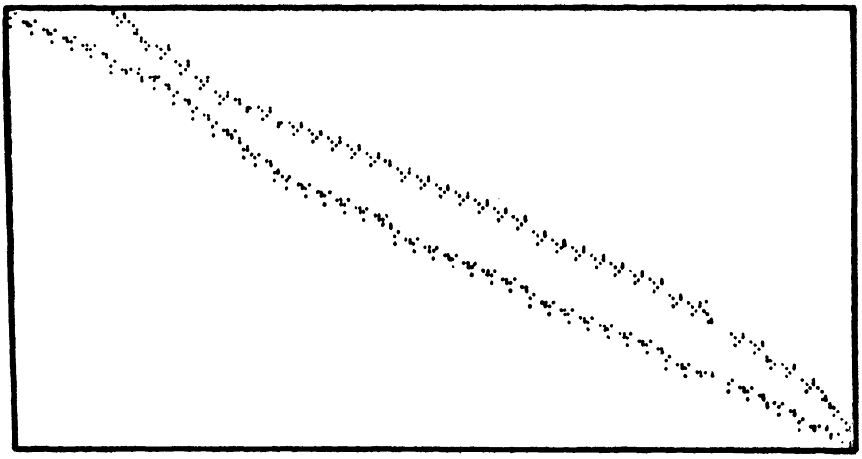


FIG. 4. *Sparsity pattern of equilibrium matrix E for Fig.* 3.

earlier, that the $m \times n$ equilibrium matrix E is assembled into $E = [E_1 E_2]$, with E_1 $m \times m$ nonsingular and banded. Moreover, the element flexibility matrix F is block diagonal with positive definite diagonal blocks and can be partitioned conformally into

$$F = \begin{bmatrix} F_1 & 0 \\ 0 & F_2 \end{bmatrix}, \quad F_1 = \operatorname{diag}(F_{11}, \cdots, F_{ss}), \quad F_2 = \operatorname{diag}(F_{s+1,s+1}, \cdots, F_{qq}),$$

where F_1 is $m \times m$ and q is the number of elements in the finite element model. Finally, f is partitioned conformally with the diagonal blocks F_{ii} of F; namely,

$$f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \quad f_1 \text{ an } m\text{-vector}, \quad f_1 = \begin{bmatrix} f_{11} \\ \vdots \\ f_{1s} \end{bmatrix}, \quad f_2 = \begin{bmatrix} f_{2,s+1} \\ \vdots \\ f_{2q} \end{bmatrix},$$

where f_{ij} is a vector of the same dimension as F_{jj}, for $j = 1, 2$.

ALGORITHM 2. *Parallel solution of the structural analysis equations* (3.4) *by the block cyclic method.*

Step 1. Compute the Cholesky factors of the diagonal blocks F_{jj} of $F_2, j = s+1, \cdots, q$ (concurrently) and compute the banded triangular factors of E_1.

Step 2. Estimate $\rho = \rho(J)$ given by (2.8) and the optimum SOR parameter ω_b given by (2.10), using the factorization obtained in Step 1.

Step 3. Choose initial approximations

$$f^{(0)} = \begin{bmatrix} f_1^{(0)} \\ f_2^{(0)} \end{bmatrix},$$

where $f_1^{(0)}$ is an m-vector, to the system force vector f and $r^{(0)}$ to the nodal displacement vector r.

Step 4. FOR $k = 0$ STEP 1 UNTIL convergence DO

1) FOR $j = s+1$ STEP 1 TO q solve (concurrently)

$$F_{jj} f_{2j}^{(k+1)} = (1 - \omega_b) F_{jj} f_{2j}^{(k)} - \omega_b E_2^T r^{(k)}.$$

2) Solve

$$E_1 f_1^{(k+1)} = (1-\omega_b)E_1 f_1^{(k)} + E_2[(1-\omega_b)f_2^{(k)} - f_2^{(k+1)}] + \omega_b p.$$

3) FOR $j=1$ STEP 1 TO s compute (concurrently) vectors

$$y_j = \omega_b F_{jj} f_{1j}^{(k)}$$

and set $y = \begin{bmatrix} y_1 \\ \vdots \\ y_s \end{bmatrix}$.

4) Solve

$$E_1^T r^{(k+1)} = (\omega_b - 1)E_1^T r^{(k)} + y.$$

Observe first that, as promised, parts of Algorithm 2 lend themselves to parallel processing. The computations of the Cholesky factors of the F_{2j}, $j = s+1, \cdots, q$, in Step 1 can be done in parallel in one major time step with $q-s$ processors. In addition, parallel algorithms are available for the triangular factorization of the banded matrix E_1 in Step 1 (see Dongarra and Sameh [1984]). Next, part 1 of Step 4 is essentially a block Jacobi iterative scheme and can be accomplished in one major time step, if $q-s$ processors are employed. Similarly, part 3 of Step 4 can also be accomplished in one major time step, if s processors are employed. Generally speaking, all parts of Step 4 involve sums of matrix-vector products and can be efficiently implemented using SAXPY or GAXPY type algorithms on pipelined vector machines (see Dongarra, Gustavson and Karp [1984] or Ortega and Voigt [1985]).

Another situation of interest in Algorithm 2 is where E is fixed and only the F_{jj} vary in multiple redesign problems. Here the original factorization of E_1 is used in each major redesign step, i.e., in each application of Algorithm 2. Furthermore, the iterative nature of the algorithm facilitates the use of the previous approximation solution $[{}^f_r]$ to (3.4) as the initial starting vectors $f^{(0)}$ and $r^{(0)}$ in the next redesign step. It should also be pointed out that the scheme described here iterates relative to the system force vector f *and* the nodal displacement vector r simultaneously; whereas, the force method determines f directly and then r and the displacement method determines r directly and then f.

A numerical comparison of the 2-block SOR parallel iterative scheme suggested here with recent direct serial and parallel algorithms for structural analysis computations is in preparation. Comparisons are being made on a variety of structural analysis problems, including two- and three-dimensional frames, plane stress, plate bending and mixed finite element problems considered in Berry, Heath, Kaneko, Lawo, Plemmons and Ward [1985] and Berry and Plemmons [1985a, b]. A description of this work will appear elsewhere.

REFERENCES

[1] O. Axelsson [1984], *Numerical algorithms for indefinite problems*, in Elliptic Problem Solvers II, Academic Press, New York.

[2] M. Berry and R. Plemmons [1985a], *Parallel algorithms for finite element structural analysis on the* HEP *multiprocessor*, Proc. of the Denelcor Workshop on the HEP, Univ. Oklahoma, March, 1985, pp. 157–180.

[3] ——— [1985b], *Computing a banded basis of the null space on the* HEP *multiprocessor*, AMS Contemporary Mathematics, to appear.

[4] M. Berry, M. Heath, I. Kaneko, M. Lawo, R. Plemmons and R. Ward [1985], *An algorithm to compute a sparse basis of the null space*, Numer. Math., to appear.

[5] Y. CHEN [1975], *Iterative methods for linear least squares problems*, Rept. CS-75-04, Dept. computer Science, Univ. Waterloo, Waterloo, Ontario, Canada.

[6] J. DE PILLIS, W. NIETHAMMER AND T. VARGA [1984], *Convergence of block iterative methods applied to sparse least squares problems*, Linear Algebra Appl., 58, pp. 327-342.

[7] J. DONGARRA, F. GUSTAVSON AND A. KARP [1984], *Implementing linear algebra algorithms for dense matrices on a vector pipelined machine*, SIAM Rev., 26, pp. 91-112.

[8] J. DONGARRA AND A. SAMEH [1984], *On some parallel banded system solvers*, Mathematics and Computer Science Div. Tech. Rept., 84-27, Argonne National Laboratory, Argonne National Laboratory, Argonne, IL.

[9] I. DUFF AND J. REID [1979], *Some design features of a sparse matrix code*, ACM Trans. Math. Software, 2, pp. 18-35.

[10] N. DYN AND W. FERGUSON, JR. [1983], *The numerical solution of equality-constrained quadratic programming problems*, Math. Comp., 41, pp. 165-170.

[11] P. GILL AND W. MURRAY [1974], *Numerical Methods for Constrained Optimization*, Academic Press, New York.

[12] G. HADLEY [1964], *Nonlinear and Dynamic Programming*, Addison-Wesley, Reading, MA.

[13] M. HEATH, R. PLEMMONS AND R. WARD [1984], *Sparse orthogonal schemes for structural optimization using the force method*, SIAM J. Sci. Stat. Comp., 5, pp. 514-532.

[14] R. HUANG [1983], *On the determination of iteration parameters for complex SOR and Chebyshev methods*, CNA Rept., 187, Univ. Texas, Austin.

[15] I. KANEKO AND R. PLEMMONS [1984], *Minimum norm solutions to linear elastic analysis problems*, Internat. J. Numer. Meth. Engrg., 20, pp. 983-998.

[16] I. KANEKO, M. LAWO AND G. THIERAUF [1982], *On computational procedures for the force method*, Internat. J. Num. Meth. Engrg., 18, pp. 1469-1495.

[17] T. MARKHAM, M. NEUMANN AND R. PLEMMONS [1985], *Convergence of a direct-iterative method for large-scale least squares problems*, Linear Algebra Appl., 69, pp. 155-167.

[18] J. ORTEGA AND R. VOIGT [1985], *Solution of partial differential equations on vector and parallel computers*, ICASE Rept., 85-1, NASA Langley, Hampton, VA; SIAM Rev., 27 (1985), pp. 169-241.

[19] R. PLEMMONS [1979], *Adjustment by least squares in geodesy using block iterative methods for sparse matrices*, Proc. U.S. Army Numerical Analysis and Computers Conference, El Paso, TX, pp. 151-186.

[20] J. ROBINSON [1973], *Integrated Theory of Finite Element Methods*, John Wiley, New York.

[21] R. VARGA [1962], *Matrix Iterative Analysis*, Prentice-Hall, Englewood-Cliffs, NJ.

[22] D. YOUNG [1971], *Iterative Solution of Large Linear Systems*, Academic Press, New York.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

A DYNAMIC PROGRAMMING ALGORITHM FOR COVERING PROBLEMS WITH (GREEDY) TOTALLY BALANCED CONSTRAINT MATRICES*

MARTIN W. BROIN† AND TIMOTHY J. LOWE‡

Abstract. Given an $m \times n$ (0, 1) matrix A possessing special structure, we consider the Minimum Cost Maximal Covering Problem (MCP):

$$\text{Min}\,\{cx+dy \mid Ax+y \geqq 1^m,\ 1^n x \leqq p;\ x \in \{0,1\}^n,\ y \in \{0,1\}^m\}.$$

We give an $O(p^2 n \min\{m^2, n^2\})$ dynamic programming algorithm for solving (MCP) when the matrix A is totally balanced. Totally balanced matrices arise naturally in tree network location problems; however, the class of totally balanced matrices is not equivalent to the class of covering matrices arising from tree network location problems: we give an example of a totally balanced matrix for which there exists no corresponding tree network covering problem.

Key words. dynamic programming, integer programming, graphs and matrices

AMS(MOS) subject classifications. 90C10, 90C39, 05C50

1. Totally balanced covering problems.

Berge [1] and Fulkerson et al. [2] provide the foundations for the study of balanced matrices. A (0, 1)-matrix A is said to be *balanced* if it does not contain an odd square submatrix of size at least three which has no identical columns and with all of its row and column sums equal to two. Fulkerson provides the following lemma:

LEMMA 1. *If an $m \times n$ (0, 1)-matrix A is balanced and if*

$$P = \{x \mid Ax \geqq 1^m,\ x \geqq 0^n\}$$

is not empty, then every vertex of this polytope has coordinates 0 *or* 1.

The *Minimum Cost Covering Problem* is

$$\text{(CP)} \qquad \text{Min}\,\{cx \mid Ax \geqq 1^m,\ x \in \{0,1\}^n\},$$

where A is an $m \times n$ (0, 1)-matrix and $c \in R^n_+$. Thus the minimum cost covering problem can be solved by relaxing the binary constraints and utilizing the linear programming relaxation as long as the covering matrix A is balanced.

A (0, 1)-matrix A is said to be *totally balanced* if it does not contain a square submatrix which has no identical columns and with all row and column sums equal to two. Clearly a totally balanced matrix is balanced and so enjoys the integer extreme point property implied by the above lemma.

Example. Let $T=(V, E)$ be a tree with vertex set V and edge set E. Each edge $e \in E$ is assumed to have a positive length associated with it. Let $X=\{x_1, \cdots, x_n\}$ and $Y=\{y_1, \cdots, y_m\}$ be two sets of points on the tree, and each point $x_j \in X$ has a nonnegative number r_j (called a radius) associated with it. The distance between two points z and z' on T, denoted $d(z, z')$ is defined to be the length of the shortest path connecting them. Let $A=(a_{ij})$ be the $m \times n$ (0, 1)-matrix defined by $a_{ij}=1$ if $d(x_j, y_i) \leqq r_j$ and $a_{ij}=0$ otherwise. Giles [3] proved that the matrix A is totally balanced. Tamir

* Received by the editors August 18, 1983, and in revised form June 17, 1985. This research was supported in part by the National Science Foundation under grant ECS-8317026.

† School of Management, University of Texas-Dallas, Richardson, Texas 75080.

‡ Krannert Graduate School of Management, Purdue University, West Lafayette, Indiana 47907. Present address, Industrial Engineering Systems Department, University of Florida, Gainesville, Florida 32611.

[11] generalized Giles' results to obtain: Let r'_i be a nonnegative radius associated with $y_i \in Y$ and define $B = (b_{ij})$ by $b_{ij} = 1$ if $d(x_j, y_i) \leqq r_j + r'_i$ and $b_{ij} = 0$ otherwise. Then the (0, 1)-matrix B is totally balanced. The matrix B is called the intersection matrix of neighborhood subtrees versus neighborhood subtrees on T. Tamir exploits this generalization, and Lemma 1, to verify the use of linear programming to solve various versions of problem (CP) on a tree network. In addition, Tamir provides a dynamic programming algorithm for the problem which directly utilizes the tree network.

The *Minimum Cost Partial Covering Problem* is

$$\text{(PCP)} \qquad \text{Min}\,\{cx + dy \mid Ax + y \geqq 1^m,\ x \in \{0,1\}^n,\ y \in \{0,1\}^m\},$$

where A is an $m \times n$ (0, 1)-matrix, $c \in R^n_+$ and $d \in R^m_+$. It can be shown that if A is totally balanced, then the matrix $[A, I]$ is totally balanced (where I is the identity matrix) and so problem (PCP) defined on a tree network can be solved by relaxing the integer conditions and using linear programming.

Kolen [7] provides an elegant $O(nm)$ algorithm to solve problem (CP) when the matrix A is totally balanced. The algorithm first solves the dual of the linear programming relaxation of (CP) via a greedy algorithm and then recovers the optimal primal solution to (CP) using complementary slackness. The algorithm requires that the matrix A be in *standard greedy* form. A (0, 1)-matrix A is standard greedy if $a_{ik} = a_{il} = a_{jk} = 1$ implies that $a_{jl} = 1$ for all i, j, k, l with $i < j$, $k < l$. In other words, the ordered matrix A does *not* contain the ordered submatrix:

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

Kolen also shows that an $m \times n$ (0, 1)-matrix is totally balanced if and only if it can be transformed by row and column permutations into standard greedy form. He also provides an $O(n^2m)$ algorithm for transforming a totally balanced matrix into standard greedy form.

Hoffman, Kolen and Sakarovitch [4] extend Kolen's algorithm to problem (PCP), by giving an algorithm for a generalization of problem (PCP) with an ordered integer m-vector b on the right-hand side of the covering constraints. Again it is assumed that the totally balanced matrix A is in standard greedy form, and an $O(mn)$ algorithm is given which solves the dual problem via a greedy procedure and recovers the optimal primal solution.

In [8], Kolen considers problems (CP) and (PCP) and the closely related plant location problem (see Kolen and Tamir [9], for example) when these problems are defined on a tree network with m nodes, each of which is a demand point. Kolen shows the relationship between these problems and gives polynomial algorithms for each problem. By exploiting properties of the tree network, an $O(m^2)$ algorithm is given to transform the matrix A into standard greedy form. Then, the $O(nm)$ algorithm in [4] is used to solve (CP) or (PCP).

For a thorough discussion of covering problems on tree networks, see [9].

The *Minimum Cost Maximal Covering Problem* is

$$\text{(MCP)} \qquad \text{Min}\,\{cx + dy \mid Ax + y \geqq 1^m,\ 1^n x \leqq p,\ x \in \{0,1\}^n,\ y \in \{0,1\}^m\},$$

where A is an $m \times n$ (0, 1)-matrix, $c \in R^n_+$, $d \in R^m_+$ and p is a positive integer. The Minimum Cost Maximal Covering Problem cannot always be solved by relaxing the integer constraints even when the covering matrix A is standard greedy. As an example, with $c=(0,0,0,0)$, $d=(2,3,4,5,5,5)$, $p=2$, and

$$A=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\1&0&0&1\\0&1&0&1\\0&0&1&1\end{bmatrix},$$

the optimal answer is $x_3=x_4=1$, $x_1=x_2=0$, $y_1=y_2=1$ and $y_3=\cdots=y_6=0$ with an optimal objective value of 5. However, if the integer constraints are relaxed, then the solution $x_1=x_2=x_3=x_4=\frac{1}{2}$, $y_1=y_2=y_3=\frac{1}{2}$ and $y_4=y_5=y_6=0$ is feasible and has an objective value of $4\frac{1}{2}$.

In the next section we give an $O(p^2 n \min\{m^2, n^2\})$ dynamic programming algorithm to solve (MCP) in the case that A is totally balanced. The algorithm works directly with the matrix A, once in standard greedy form.

Megiddo et al. [10] consider a maximum coverage facility location problem on a tree network, where if demand point i on the tree is "covered" by a facility, then a gain of $w_i>0$ is obtained. Demand point i has an associated cover radius, $r_i \geqq 0$ and is covered if a facility location x_j is chosen where $d(y_i, x_j) \leqq r_i$. In the problem, there is an upper bound p on the number of new facility locations that can be chosen. Clearly, the maximum coverage problem can be formulated in terms of (MCP), with $c=0$, $d_i=w_i$, $i=1,\cdots,m$, and taking A as the intersection matrix of demand point neighborhoods and potential facility locations. In [10] a dynamic programming type algorithm is given which works directly with, and exploits properties of, the tree network. Also, Hsu [5] considers a version of the maximum coverage problem on a tree network and gives an algorithm which works directly with the tree network.

Prior to giving our algorithm, we first show that it can be applied to a broader class of problems than "covering problems" on a tree network. We demonstrate this by showing that for a particular totally balanced matrix, there is no tree network T for which matrix is the intersection matrix of neighborhood subtrees versus neighborhood subtrees on T. For our result, we need the following lemma.

LEMMA 2. *Let T be a tree with vertex set V and arc set E, with nonnegative arc lengths. Let $\{a,b,c,d\}$ be a subset of V, and let r_v denote the nonnegative radius associated with each $v\in V$. Further, let P_{uw} be the (undirected) shortest path between points u and w in T. If there exist points x_1 and x_2 on T, with associated nonnegative radii r_1 and r_2 such that*

$$\{a,b\}=V_1=\{v\in\{a,b,c,d\}\,|\,d(x_1,v)\leqq r_1+r_v\}, \quad \textit{and}$$
$$\{c,d\}=V_2=\{v\in\{a,b,c,d\}\,|\,d(x_2,v)\leqq r_2+r_v\}, \tag{1}$$

then $P_{ab}\cap P_{cd}=\varnothing$.

Proof. Assume that (1) holds and that there exists $z\in T$ such that $z\in P_{ab}\cap P_{cd}$. Due to properties of a tree, it follows that either (i) $z\in P_{ad}$ and $z\in P_{cb}$, or (ii) $z\in P_{ac}$ and $z\in P_{db}$, or that both (i) and (ii) hold. Without loss of generality, we assume that (i) holds. For the remainder of the proof, it may be helpful to refer to Fig. 1, which provides a representation of P_{ab}, P_{cd} and z. In the figure, f is the closest point in P_{ab} to c and g is the closest point in P_{cd} to b.

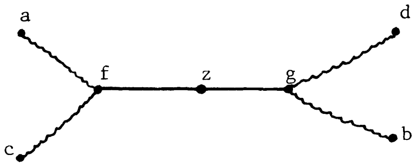


FIG. 1

The purpose of the figure is to show the orientation of a, b, c, and d relative to z. We note that, for example, c may be on P_{ab} in which case c coincides with f. Also, for example, if P_{ab} intersects P_{cd} at a single point (z), then both f and g coincide with z (in which case, both (i) and (ii) above hold).

Given that (i) holds, from the properties of a tree, we note that for any $u \in P_{az} \cup P_{cz}$ and $w \in P_{dz} \cup P_{bz}$, $d(u, w) = d(u, z) + d(z, w)$. Let x_j' be the closest point in $P_{ab} \cup P_{cd}$ to $x_j, j = 1, 2$. Again from the properties of a tree, for any $u \in P_{ab} \cup P_{cd}$, $d(u, x_j) = d(u, x_j') + d(x_j', x_j), j = 1, 2$.

First, suppose that $x_1' \in P_{az} \cup P_{cz}$. Since $b \in V_1$ but $d \notin V_1$, we have $d(b, x_1) - r_b \leqq r_1 < d(d, x_1) - r_d$. But,

$$\begin{aligned} d(b, x_1) - d(d, x_1) &= d(b, z) + d(z, x_1') + d(x_1', x_1) - (d(d, z) + d(z, x_1') + d(x_1', x_1)) \\ &= d(b, z) - d(d, z) \end{aligned}$$

so that

$$d(b, z) - r_b < d(d, z) - r_d. \tag{2}$$

Since $b \in V_1$ but $c \notin V_1$, we have $d(b, x_1) - r_b \leqq r_1 < d(c, x_1) - r_c$. But $d(b, x_1) = d(b, z) + d(z, x_1') + d(x_1', x_1)$. Also, due to the triangle inequality, $d(c, x_1) \leqq d(c, z) + d(z, x_1') + d(x_1', x_1)$ so that $d(b, x_1) - d(c, x_1) \geqq d(b, z) - d(c, z)$. Thus,

$$d(b, z) - r_b < d(c, z) - r_c. \tag{3}$$

If $x_2' \in P_{az} \cup P_{cz}$, then with $d \in V_2$ but $b \notin V_2$, we have $d(d, x_2) - r_d \leqq r_2 < d(b, x_2) - r_b$. But $d(d, x_2) - d(b, x_2) = d(d, z) + d(z, x_2') + d(x_2', x_2) - (d(b, z) + d(z, x_2') + d(x_2', x_2)) = d(d, z) - d(b, z)$ so that $d(d, z) - r_d < d(b, z) - r_b$ which contradicts (2). If $x_2' \in P_{dz} \cup P_{bz}$, then with $c \in V_2$ but $b \notin V_2$ we have $d(c, x_2) - r_c \leqq r_2 < d(b, x_2) - r_b$. But $d(c, x_2) = d(c, z) + d(z, x_2') + d(x_2', x_2)$ and $d(b, x_2) \leqq d(b, z) + d(z, x_2') + d(x_2', x_2)$ so that $d(c, x_2) - d(b, x_2) \geqq d(c, z) - d(b, z)$. Thus, $d(c, z) - r_c < d(b, z) - r_b$ which contradicts (3).

Due to symmetry, similar contradictions can be established when $x_1' \in P_{dz} \cup P_{bz}$.

Using Lemma 2, we can now prove the following.

PROPOSITION. *For the totally balanced matrix*

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix},$$

there exists no tree, T, where A is the intersection matrix of neighborhood subtrees versus neighborhood subtrees on T.

Proof. We assume a tree, T, exists and show a contradiction. For the matrix A, let the row indices correspond to the indices of a subset of the vertex set of T and let the column indices correspond to the index set of points on the tree. Using Lemma 2 with $a=1$, $b=5$, $c=2$ and $d=6$ and the first and third columns of A, it follows that $P_{15}\cap P_{26}=\emptyset$. Using Lemma 2 with $a=1$, $b=5$, $c=4$ and $d=6$ and the first and fourth columns of A, we have $P_{15}\cap P_{46}=\emptyset$. Thus, $P_{15}\cap\{P_{26}\cup P_{46}\}=\emptyset$.

Using Lemma 2 with $a=3$, $b=5$, $c=2$ and $d=6$ and the second and third columns of A, we have $P_{35}\cap P_{26}=\emptyset$. Using Lemma 2 with $a=3$, $b=5$, $c=4$, and $d=6$ and the second and fourth columns of A, we have $P_{35}\cap P_{46}=\emptyset$.

It now follows that $\{P_{15}\cup P_{35}\}\cap\{P_{26}\cup P_{46}\}=\emptyset$. Since $\{P_{15}\cup P_{35}\}$ and $\{P_{26}\cup P_{46}\}$ are disjoint subtrees of T, there exists some point z in the shortest path connecting these subtrees such that $z\in P_{uw}$ for all $u\in\{P_{15}\cup P_{35}\}$ and all $w\in\{P_{26}\cup P_{46}\}$. Thus $z\in P_{12}\cap P_{34}$ so that $P_{12}\cap P_{34}\neq\emptyset$. But then, letting $a=1$, $b=2$, $c=3$ and $d=4$, and using the fifth and sixth columns of A, we have a contradiction via Lemma 2.

2. A dynamic programming algorithm for (MCP). A row and column index pair (i,j) is called a *root* of a standard greedy (0, 1)-matrix A if $a_{ij}=1$ and it is the last nonzero entry in the jth column and the last nonzero entry in the ith row of A. An essential property exploited by the algorithm is that if a standard greedy matrix has more than one root, then (MCP) can be decomposed into as many subproblems as there are roots and the only constraint linking the subproblems together is the constraint $1^n x\leqq p$.

In what follows, we consider an equivalent formulation of (MCP):

$$\begin{aligned}
&\text{Max} -\sum_{j=1}^{n} c_j x_j+\sum_{i=1}^{m} d_i y_i\\
&\text{s.t.}\quad \sum_{j=1}^{n} a_{ij}x_j-y_i\geqq 0,\qquad i=1,\cdots,m,\\
&\qquad\ \sum_{j=1}^{n} x_j\leqq p,\\
&\qquad\ x_j,\ y_i\in\{0,1\},\quad j=1,\cdots,n;\quad i=1,\cdots,m.
\end{aligned}\tag{4}$$

We define a *subproblem* of (MCP) as problem (4) restricted to a subset I of rows of A and a subset J of columns of A, and nonnegative integer $\pi\leqq p$ as the constraint on $\sum_{j\in J} x_j$; we write the optimal objective function value of the subproblem as $V(I, J, \pi)$. Two subproblems defined by (I_r, J_r, π_r) and (I_s, J_s, π_s) are said to be *independent* if $I_r\cap I_s=\emptyset$, $J_r\cap J_s=\emptyset$ and $a_{ij}=0$ for every $i\in I_r$, $j\in J_s$, and $i\in J_s$, $j\in J_r$.

THEOREM. *If a subproblem defined by π and index sets I and J has K roots, then the subproblem can be decomposed in K mutually independent subproblems* (*one for each root*) *such that*

$$V(I, J, \pi)=\max\left\{\sum_{k=1}^{K} V(I_k, J_k, \pi_k)\,\middle|\,\sum_{k=1}^{K}\pi_k\leqq\pi,\ \pi_k\geqq 0,\ \text{integer},\ k=1,\cdots,K\right\}.$$

To prove the theorem, we will show that any column (or row) of the submatrix defining the subproblem can be associated with exactly one of the K roots. For a nonzero entry (i',j) in a standard greedy matrix, we define the last nonzero entry in row i and the last nonzero entry in column j to be *feasible destinations* for the given entry (i,j). By definition, any root is its own unique feasible destination.

For a nonzero entry (i, j), a *feasible walk* (from (i, j)) is a sequence of distinct nonzero entries $(i_t, j_t), t = 1, \cdots, T$ such that $(i_1, j_1) = (i, j), (i_T, j_T)$ is a root, and (i_{t+1}, j_{t+1}) is a feasible destination of (i_t, j_t), $t = 1, \cdots, T-1$. If (i, j) is a root, then the feasible walk is just $(i_1, j_1) = (i, j)$. Given a feasible walk (i_j, j_t), $t = 1, \cdots, T$, where $T > 1$, we say that nonzero entry (i', j') is on the *feasible walk*, if either

$$\text{(i)}\ i_t = i' = i_{t+1} \quad \text{and}\ j_t \leqq j' \leqq j_{t+1}, \quad \text{for some } t, \quad 1 \leqq t \leqq T-1,$$

or

$$\text{(ii)}\ j_t = j' = j_{t+1} \quad \text{and} \quad i_t \leqq i' \leqq i_{t+1}, \quad \text{for some } t, \quad 1 \leqq t \leqq T-1.$$

The next property follows immediately from the definitions of feasible destinations and feasible walks.

Property 1. Let (i_t, j_t), $t = 1, \cdots, T$, be a feasible walk from nonzero entry (i, j). Further, suppose nonzero entry (i', j') is on the feasible walk where we suppose, without loss of generality, that (i) above holds at some value of t. Then (i', j'), $(i_{t+1}, j_{t+1}), \cdots, (i_T, j_T)$ is a feasible walk from (i', j'), when $j' < j_{t+1}$. Otherwise, when $j' = j_{t+1}$, then $(i_{t+1}, j_{t+1}), \cdots, (i_T, j_T)$ is a feasible walk from (i', j').

LEMMA 3. *Two feasible walks starting at the same nonzero entry* $(\hat{i}, \hat{j})$ *terminate at the same root.*

Proof. If nonzero entry $(\hat{i}, \hat{j})$ is in the last row or last column, the result is true since there is only one distinct feasible walk from $(\hat{i}, \hat{j})$. Thus consider a nonzero entry $(\hat{i}, \hat{j})$ which is not in the last row or column. Further, suppose the result is true for all nonzero entries (i, j) where either $i \geqq \hat{i}$ and $j > \hat{j}$, or $i > \hat{i}$ and $j \geqq \hat{j}$.

If $(\hat{i}, \hat{j})$ is a root, then the result is true by definition of a feasible walk from a root. Thus we suppose that $(\hat{i}, \hat{j})$ is not a root and consider the case where $(\hat{i}, \hat{j})$ has only one feasible destination. Without loss of generality suppose this feasible destination is $(\hat{i}, j)$ with $j > \hat{j}$. Clearly, $(\hat{i}, j)$ is on any feasible walk from $(\hat{i}, \hat{j})$. But then using Property 1 and the induction hypothesis for entry $(\hat{i}, j)$ proves the result.

Now consider the case where $(\hat{i}, \hat{j})$ has two distinct feasible destinations $(\hat{i}, j)$ and $(i, \hat{j})$ with $j > \hat{j}$ and $i > \hat{i}$. Because the matrix is in standard greedy form, it follows that entry (i, j) is nonzero. If (i, j) is a root, the result is true. Otherwise, using Property 1 and the induction hypothesis on (i, j) proves the result.

From Lemma 3, we have that any nonzero entry is associated (through the definition of a feasible walk) with exactly one root. Furthermore, from Property 1, it follows that all nonzero entries in a given row (or column) are associated with the same root. This observation proves the Theorem.

Using the concept of a feasible walk, we have identified the root with which a given nonzero entry is associated. By essentially "reversing" the walk from a root, we can easily find the subproblem rows and columns associated with the given root. Henceforth, we will identify a subproblem with a single root by the root itself.

There are two major procedures used in the dynamic programming algorithm to solve (MCP). These two procedures recursively call each other in a well defined way.

The first procedure, denoted as SUBPROB $((i, j), \pi)$ solves a subproblem of (MCP), with a *single* root (i, j) by setting at most $\pi (\leqq p)$ x's in the subproblem equal to one. In SUBPROB $((i, j), \pi)$, we check to see if x_j should be set to one or zero. When x_j is set to one, we remove column j and all rows covered by column j (rows i' where $a_{i'j} = 1$). When x_j is set to zero, we remove only column j.

In either case ($x_j = 1$ or $x_j = 0$), for the remaining matrix (after column and row removal) we find the roots and solve the problem using procedure ALLOCATE $(\cdot)$.

Procedure ALLOCATE (roots, π) solves the problem of optimally allocating π over a subproblem with a set of K roots $\{(i_k, j_k), k=1, \cdots, K\}$ and is, as justified by the theorem, the one-dimensional packing problem:

$$V((i_k, j_k), k=1, \cdots, K, \pi) = \text{Max} \sum_{k=1}^{K} V((i_k, j_k), \pi_k)$$

$$\text{s.t.} \sum_{k=1}^{K} \pi_k \leqq \pi$$

$$\pi_k \geqq 0, \quad \text{integer.}$$

Since $V((i_k, j_k), \pi_k)$ is needed for various values of π_k, procedure ALLOCATE (roots, π) calls procedure SUBPROB $(\cdot)$. For completeness, the procedures are stated formally below.

Let I be a subset of row indices and J be a subset of column indices of the matrix A. Given any $j \in J$, define $I(j)$ as the index set of rows in I for which there is a one in column j; and for any $i \in I$, define $J(i)$ as the index set of columns in J for which there is a one in row i. More formally,

$$I(j) \equiv \{i \in I | a_{ij} = 1\}, \quad \text{and} \quad J(i) \equiv \{j \in J | a_{ij} = 1\}.$$

Let (i_*, j_*) be a root of the submatrix defined by I and J, and let I_* and J_* denote the subproblem associated with the root (i_*, j_*). We note that in procedure SUBPROB we do not need to explicitly compute I_* and J_*.

```
Procedure SUBPROB ($(i_*, j_*)$, $\pi$);
comment: solve a subproblem with a single root $(i_*, j_*)$ and $\pi$.
begin
  $M1 \leftarrow \{i \in I(j_*) | i \leqq i_*\}$;
  $N1 \leftarrow \{j \in U_{i \in M1} J(i) | j < j_*\}$;
  $M2 \leftarrow \{i \in U_{j \in N1} I(j)\} \backslash I(j_*)$;
  comment: Case 1: $x_{j_*}$ is set to zero, find new roots.
  Find all $\underline{\text{ROOTS}}^1$ in the submatrix defined by the index sets $I_*$ and $J_* \backslash \{j_*\}$;
  comments: All $\underline{\text{ROOTS}}^1$ will be roots of the submatrix defined by $M1$ and $N1$.
  if $|\text{ROOTS}^1| = 0$ then $Z^1 \leftarrow 0$
  else $Z^1 \leftarrow$ ALLOCATE (($\text{ROOTS}^1$), $\pi$);
  comment: Case 2: $x_{j_*}$ is set to one, find new roots.
  Find all $\underline{\text{ROOTS}}^2$ in the submatrix defined by the index sets $I_* \backslash I_*(j_*)$ and $J_* \backslash \{j_*\}$;
  comment: All $\underline{\text{ROOTS}}^2$ will be roots of the submatrix defined by $M2$ and $N1$.
  if ($|\text{ROOTS}^2| = 0$) or ($\pi = 1$) then $Z^2 \leftarrow \sum_{i \in M1} d_i - c_{j_*}$
  else $Z^2 \leftarrow \sum_{i \in M1} d_i - c_{j_*} +$ ALLOCATE (($\text{ROOTS}^2$), $\pi - 1$);
  comment: Choose the maximum of Case 1 and Case 2.
  return $V((i_*, j_*), \pi) \leftarrow \max\{Z^1, Z^2\}$
end
```

```
Procedure ALLOCATE ((roots), $\pi$):
begin
  comment: $K$ is the number of roots in the subproblem.
  $K \leftarrow |\text{roots}|$;
```

```
    comment: Let $(i_k, j_k)$ denote the $k$th root.
        $V((\text{roots}), \pi) \leftarrow \text{Max} \sum_{k=1}^{K} \text{SUBPROB}\,(i_k, j_k), \pi_k)$
            s.t. $\sum_{k=1}^{K} \pi_k \leqq \pi$
                $\pi \geqq 0$, integer;
    return $V((\text{roots}), \pi)$
end
```

We remark that in Case 1 of SUBPROB (·), the index sets $M1$ and $N1$ are always subsets, and may in fact be proper subsets, of I_* and $J_*\backslash\{j_*\}$, respectively. However, it is easy to verify that any roots of the submatrix defined by index sets I_* and $J_*\backslash\{j_*\}$ will be roots of the submatrix defined by $M1$ and $N1$. Similarly, in Case 2 it is easily verified that any roots of the submatrix defined by the index sets $I_*\backslash I_*(j_*)$ and $J_*\backslash\{j_*\}$ will be roots of the submatrix defined by $M2$ and $N1$.

In procedure ALLOCATE, we are solving a one-dimensional packing problem with separable objective functions. Each of these objective functions is the solution $V((i_k, j_k), \pi_k)$ of a subproblem with a single root (i_k, j_k) and π_k. We note that *each* time ALLOCATE is accessed, there are no more than $\min\{m, n\}$ such objective functions, i.e., $K \leqq \min\{m, n\}$. Further, it is always the case that $\pi \leqq p$.

Karush [6] gives an $O(p^2K)$ dynamic programming algorithm for solving the one-dimensional packing problem with K objective functions and p the upper bound on the packing constraint. Thus, it follows that ALLOCATE can be solved in $O(p^2 \min\{m, n\})$. In fact, Karush's algorithm can simultaneously solve the one-dimensional packing problem for all $\pi = 1, 2, \cdots, p$ with no increase in computational effort.

We now consider the total number of roots that will be identified in the solution of (MCP). Consider column $n-l+1$ (which is the lth column of A measured from the right). Certainly the bottom-most entry of this column will be a root since the algorithm will eventually set all $x_j, j > n-l+1$, to zero at some stage. Let i be a row index, where $a_{i,n-l+1} = 1$, and this entry is not the last nonzero entry of column $n-l+1$. Further, let i' denote the row index, where entry $(i', n-l+1)$ is the first nonzero entry in column $n-l+1$ below row i. If entry $(i, n-l+1)$ is ever a root, there must exist a column $q(i)$ of A, $q(i) > n-l+1$, where $a_{i,q(i)} = 0$ and $a_{i',q(i)} = 1$. (If there were no such column $q(i)$, then every subproblem containing entry $(i, n-l+1)$ would also contain entry $(i', n-l+1)$, in which case $(i, n-l+1)$ could not be a root). But then since A is a greedy matrix, it follows that for every $r \geqq i'$, if $a_{r,n-l+1} = 1$, then $a_{r,q(i)} = 1$. From this it follows that for each nonzero entry $a_{t,n-l+1}$ of column $n-l+1$, which is ever identified as a root, there must be a distinct column $q(t)$ to the right of column $n-l+1$. Thus, it now follows that there are at *most* l potential roots in column $n-l+1$, and from this it follows that an upper bound on the number of distinct roots of A identified in the solution of (MCP) is $n \min\{m, n\}$.

We now consider the number of one-dimensional packing problems that must be solved via ALLOCATE. For any given roots (i, j) and value of $\pi \leqq p$, SUBPROB calls ALLOCATE with parameters (roots^1, π) and (roots, $\pi - 1$). We note that for a fixed root (i, j), roots^1 and roots^2 are invariant with respect to π. Further, as noted earlier, ALLOCATE solves the packing problem for *all* values of $\pi = 1, \cdots, p$, simultaneously. Thus, even though (for a fixed root) ALLOCATE may be accessed several times as π

varies, it only needs to be solved twice (with roots[1] and roots[2]) for any root and the results stored for later use.

Finally, we note that a set of roots for any subproblem can be found in $O(mn)$ using the concept of a feasible walk. It now follows that (MCP) can be solved in $O(n \min\{m, n\}(mn + p^2 \min\{m, n\})) = O(p^2 n \min\{m^2, n^2\})$, when $p^2 \geqq \max\{m, n\}$.

The algorithm given in [10] to solve the maximum coverage location problem on a tree network is reported to be of $O(n^2 p)$, where n is the number of potential facility locations. This reported complexity bound appears to be dependent upon concavity of the gain function (as a function of the number of locations chosen) over any subtree of the tree network. That the gain function is not always concave is evident from Fig. 2, a tree with 7 vertices. In the figure, each vertex is both a demand point and a potential facility location, each arc is of length 1, all w_i are 1, and all r_i are one. Letting $f(k)$ be the maximal gain that can be attained by placing at most k facilities, it is easy to see that $f(1) = 4$, $f(2) = 5$, and $f(3) = 7$. But then $2f(2) < f(1) + f(3)$ so that f is not concave.

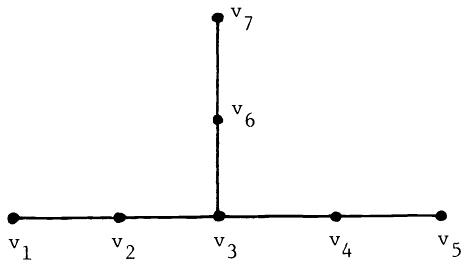


FIG. 2

We wish to point out that we were guilty of a similar error in an early version of this paper, where we falsely asserted that the optimal value of SUBPROB $((i, j), \pi)$ was concave in π. After our error was pointed out by the referees, we discovered the error in [10]. We also wish to note that although our algorithm is applicable to a broader class of problems than those addressed in [10], the preferable approach to maximum coverage problems on trees may be that of [10] (appropriately modified), since to solve such a problem with our algorithm, it would be necessary to initially generate the matrix A.

Acknowledgments. We wish to thank Antoon Kolen of Erasmus University, Rotterdam, and Eitan Zemel of Northwestern University for their comments on an earlier version of this paper. In particular, we wish to credit Professor Kolen for providing the proof of the theorem as given in this paper. The proof herein is an improvement over the proof given in the earlier version.

REFERENCES

[1] C. BERGE, *Balanced matrices*, Math. Prog., 2 (1972), pp. 19-31.

[2] D. R. FULKERSON, A. J. HOFFMAN AND R. OPPENHEIM, *On balanced matrices. Pivoting and extensions*. Math. Prog. Study, 1 (1974), pp. 120-132.

[3] R. GILES, *A balanced hypergraph defined by subtrees of a tree*, ARS Combinatoria, 6 (1978), pp. 179-183.

[4] A. J. HOFFMAN, A. W. J. KOLEN AND M. SAKAROVITCH, *Totally balanced and greedy matrices*, this Journal, 6 (1985), pp. 721-730.

[5] W. L. HSU, *The distance-domination numbers of trees*, Oper. Res., Lett., 1 (1982), pp. 96-100.

[6] W. KARUSH, *A general algorithm for the optimal distribution of effort*, Management Sci., 9 (1962), pp. 50–72.

[7] A. W. J. KOLEN, *Location problems on trees and in the rectilinear plane*, Ph.D. Dissertation, Mathematisch Centrum, Amsterdam, 1982.

[8] ———, *Solving covering problems and the uncapacitated plant location problem on trees*, Eur. J. Oper. Res., 12 (1983), pp. 266–278.

[9] A. W. J. KOLEN AND A. TAMIR, *Covering problems*, in Discrete Location Theory, P. Mirchandani and R. L. Francis, eds., John Wiley, New York, to appear.

[10] N. MEGIDDO, E. ZEMEL AND S. L. HAKIMI, *The maximum coverage location problem*, this Journal, 4 (1983), pp. 253–261.

[11] A. TAMIR, *A class of balanced matrices arising from location problems*, this Journal, 4 (1983), pp. 363–370.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

003

A PROPERTY OF THE LEGENDRE DIFFERENTIAL EQUATION AND ITS DISCRETIZATION*

F. ALBERTO GRÜNBAUM†

Abstract. We observe that a most natural discretization of the Legendre differential equation reproduces *exactly* the eigenvalues of this operator. We also identify the eigenvectors.

Key words. Legendre equation, discretization, tridiagonal, Jacobi equation, Hahn polynomials

AMS(MOS) subject classifications. 65F15, 65L10, 65D99

1. Introduction. The computation of the spectrum of a singular boundary value problem of the form

$$\left(\frac{d}{dx}\left(a(x)\frac{d}{dx}\right)+b(x)\right)f(x)=\lambda c(x)f(x) \tag{1}$$

with $0<a(x)$, $-1<x<1$, $a(\pm 1)=0$, can be handled with varying degrees of sophistication.

One chooses a grid of points $(x_1, \cdots, x_n)$, replaces the function $f(x)$ by the vector $(f(x_1), \cdots, f(x_n))^T$ and the differential operator by a tridiagonal (or higher order) matrix whose entries are obtained by sampling $a(x)$, $b(x)$, $c(x)$ at points related to the grid. Any such discretization introduces obvious errors, the most blatant being that (1) has, under appropriate conditions on $a(x)$, $b(x)$, $c(x)$ an infinite discrete spectrum of simple eigenvalues

$$0>\lambda_1>\lambda_2>\lambda_3>\cdots>\lambda_n>\cdots$$

while the matrix in (2) below has a simple but finite spectrum given by

$$\mu_1>\mu_2>\cdots>\mu_n.$$

The most that one can expect is that the lowest eigenvalues λ_j be approximated by the corresponding μ_j as n grows, that is

$$\mu_j \approx \lambda_j, \qquad j \text{ small compared to } n.$$

We will use the notation trid $(p(i), q(i), r(i))$ for a matrix M with $M_{i,i-1}=p(i)$, $M_{i,i}=q(i)$, $M_{i,i+1}=r(i)$, $M_{ij}=0$ if $|i-j|>1$.

Probably the most common discretization of the operator

$$\frac{d}{dx}a(x)\frac{d}{dx}$$

is given by the matrix

$$\frac{1}{h^2}\operatorname{trid}\,(a(x_{i-1/2}), -(a(x_{i-1/2})+a(x_{i+1/2})), a(x_{i+1/2})). \tag{2}$$

See [1, p. 363].

* Received by the editors February 18, 1985, and in revised form July 17, 1985. This work was partially supported by the Applied Mathematical Sciences Subprogram, Office of Energy Research, U.S. Department of Energy under contract DE-AC03-76SF00098, by the National Science Foundation under grant DMS-84-03232, and by the Office of Naval Research under contract N00014-84-C-0159.

† Department of Mathematics, University of California, Berkeley, California 94720.

2. Legendre equation. Here we make the observation that for the Legendre equation

$$\left(\frac{d}{dx}\left((1-x^2)\frac{d}{dx}\right)\right)f(x)=\lambda f(x)$$

if one chooses

$$x_i=-1+\frac{2(i-1)}{n}, \qquad i=1,\cdots,n+1 \tag{3}$$

and instead of (2) uses the less sophisticated $n\times n$ matrix

$$\frac{n^2}{4}\cdot \operatorname{trid}\,(a(x_i),-(a(x_i)+a(x_{i+1})),a(x_{i+1})), \tag{3'}$$

one has the *remarkable* fact that for each n,

$$\mu_j=\lambda_j, \qquad j=1,\cdots,n.$$

In this case one knows that $\lambda_j=-j(j-1)$, $i=1,2,\cdots$. As an extra bonus, it turns out that for each n, (3′) is a matrix with integer entries and integer eigenvalues. If one does not insist on *normalized* eigenvectors, then (3′) also has eigenvectors with integer entries.

THEOREM. *The matrix* (3′) *given above has eigenvalues* $\mu_1,\cdots,\mu_n$ *given by*

$$\mu_j=-(j-1)j, \qquad j=1,\cdots,n$$

and normalized eigenvectors $X_{.,1}, X_{.,2},\cdots,X_{.,n}$ *with components* $X_{i,j}$ *given by*

$$X_{i,j}=P_{j-1,n-1}\left(-1+\frac{2(i-1)}{n-1}\right).$$

Here $P_{.,n-1}$ denotes a set of n polynomials known as the Gram [1, p. 114] or Hahn [2] polynomials depending on the author's background. They correspond to the choice $\alpha=\beta=0$ in the family of polynomials discussed below. The proof is given after some comments.

Remarks. (1) These polynomials were considered by Chebyshev, see [5].

(2) The Gram polynomials $P_{.,n}(x)$ are defined by the recursion relation

$$P_{j,n}(x)=c_{j-1}xP_{j-1,n}(x)-c_{j-1}/c_{j-2}P_{j-2,n}(x), \qquad 1\leqq j\leqq n$$

with

$$c_j=\frac{n}{j+1}\sqrt{\frac{4(j+1)^2-1}{(n+1)^2-(j+1)^2}}$$

and

$$P_{-1,n}(x)=0, \qquad P_{0,n}(x)=\frac{1}{\sqrt{n+1}}.$$

They are orthonormal in the set $x_i=-1+2i/n$, $i=0,\cdots,n$ with respect to the measure that assigns weight 1 to each point.

(3) The Hahn polynomials given by

$$h_j(x,\alpha,\beta,n)\equiv h_j(x)={}_3F_2(-j,-x,j+\alpha+\beta+1;-n,\alpha+1;1),$$
$$x=0,1,\cdots,n, \quad j=0,\cdots,n$$

satisfy, for each fixed j, a second order difference equation in x [2], [3], [4]. This means that the $(n+1)\times(n+1)$

$$\text{trid}\,(D(i), -(D(i)+B(i)), B(i)), \qquad 0 \leqq i \leqq n \tag{4}$$

with $D(i)=i(n+\beta-i)$, $B(i)=(n-1-i)(\alpha+1+i)$ has eigenvectors h_j with eigenvalues $-j(j+\alpha+\beta+1)$.

In the case $\alpha=\beta=0$ the Hahn polynomials are related to the Gram polynomials, $P_{j,n}(-1+2i/n)=\mu_{j,n}h_j(i)$ with $\mu_{j,n}$ constants; moreover, from the definition above it is clear that $j!\ n(n+1)\cdots(n+j-1)h_j(i)$ is an integer for each integer value of i.

These polynomials are known to be a good uniform approximation to the Legendre polynomials only for values of j small compared to $n^{1/2}$, and for large values of j the Gram polynomials have wild oscillations, see [6], [7]. An illustration of this phenomenon is given by the unnormalized eigenvector of (3′) corresponding to the largest eigenvalue, namely

$$X_{n,j}=\binom{n}{j}(-1)^j, \qquad j=0, 1, \cdots, n.$$

Proof. Use the identity

$$\frac{n^2}{4}(1-x_i^2)=(i-1)(n-i+1)$$

to conclude that (3′) is exactly the matrix appearing in the second order difference equation given above for $\alpha=\beta=0$.

Note. The simpler example of Laplace's equation $f''(x)=\lambda f$, $f(\pm 1)=0$, discretized by $(1/h^2)$ trid $(1, -2, 1)$ has a spectrum given by

$$\mu_j=-n^2\sin^2 j\pi/2(n+1), \qquad j=1,\cdots,n$$

and only for j small compared to n we get

$$\mu_j \sim -\frac{j^2\pi^2}{4}\equiv\lambda_j.$$

3. Jacobi equation. The Hahn polynomials for α, $\beta>-1$ provide natural discrete analogues for the Jacobi polynomials. They also satisfy second-order difference equations with eigenvalues exactly matching those of the differential operator [2], [3], [4]. However, it is only for $\alpha=\beta=0$ (the Legendre case) that one encounters this matrix as a "natural" discretization of the differential operator

$$\frac{1}{(1-x)^\alpha(1+x)^\beta}\frac{d}{dx}\left((1-x)^{\alpha+1}(1+x)^{\beta+1}\frac{d}{dx}\right).$$

The differential operator can be written in the nonself-adjoint form

$$(1-x^2)\frac{d^2}{dx^2}+((\beta-\alpha)-(\alpha+\beta+2)x)\frac{d}{dx} \tag{5}$$

and one can wonder if any "natural" discretization of (5) will give a matrix with eigenvalues $-j(j+\alpha+\beta+1)$. The answer is probably no; however, a look at (4) shows that the choice in (3) and the replacement of (5) by

$$(1-x^2)D_-D_+-(\beta+1)(1+x)D_-+(\alpha+1)(1-x)D_+$$

with $D_-D_+ = \text{trid}\,(1, -2, 1)$, $D_- = \text{trid}\,(-1, 1, 0)$, $D_+ = \text{trid}\,(0, -1, 1)$ gives a matrix with eigenvalues $-j(j+\alpha+\beta+1)$.

The Legendre equation is special in the sense that a very simple-minded discretization of the differential equation gives the second order difference equation satisfied by the corresponding Hahn polynomials.

Acknowledgment. It is a pleasure to thank two referees for comments and suggestions.

REFERENCES

[1] G. Dahlquist and A. Björck, *Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[2] S. Karlin and J. McGregor, *The Hahn polynomial, formulas and an application*, Scripta Math., 26 (1961), pp. 33-46.

[3] P. Lesky, *Orthogonale Polynomisysteme als Losungen Sturm-Liouvillescher Differenzengleichungen*, Monat. Math., 66 (1962), pp. 203-214.

[4] M. Perlstadt, *Chopped orthogonal polynomial expansions, some discrete cases*, this Journal, 4 (1983), pp. 94-100.

[5] G. Szegö, *Orthogonal Polynomials*, AMS Colloquium Publications 23, American Mathematical Society, Providence, RI, 1939, pp. 33-34.

[6] M. Wayne Wilson, *On the Hahn polynomials*, SIAM J. Math. Anal., 1 (1970), pp. 131-139.

[7] S. K. Zaremba, *Some properties of polynomials orthogonal over the set* $\langle 1, 2, \cdots, N\rangle$, Ann. Mate. Pura Appl. (IV) 105 (1975), pp. 333-345.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

004

PARAMETRIC LOWER BOUND FOR ON-LINE BIN-PACKING*

G. GALAMBOS†

Abstract. In this note we give lower bounds for such a one-dimensional bin-packing problem, in which we can use only one-line rules to pack the elements, and the maximal size of the elements are bounded. Our lower bound contains, as a special case, the result given by Liang [Inform. Proc. Lett., 10 (1980), pp. 76-79].

Key words. one-dimensional bin-packing, heuristic algorithm, worst-case analysis, on-line algorithms

1. Introduction. Let us consider one of the simplest draftings of the one-dimensional bin-packing problem: There is given a list $L=\{p_1, p_2, \cdots, p_m\}$, and let us suppose that an l_i size belongs to each $p_i \in L$ $(0<l_i \leqq 1)$. The problem is to pack the pieces (the elements of the list) into bins of unit capacity, while attempting to minimize the number of bins needed for packing. This problem belongs to the $\mathcal{NP}$-hard problems [2]; therefore many approximation algorithms for its solution were examined [1], [3], [5]. We will consider those algorithms only, which try to pack one piece at a time, and after having placed the piece it will be never moved again. We call these algorithms as *on-line algorithms.*

Different on-line algorithms were developed for the wide variety of applications, for example to solve the one-dimensional cutting stock problem, to allocate spaces on a disc or in a computer memory, to operate a dynamic multitasking, to solve some scheduling problems, etc.

The performance of an algorithm A is measured by its asymptotic performance ratio R_A (see below). The first lower bound for the R_A of an on-line algorithm was given by Yao [5]. He concluded that there is no such on-line algorithm, in which $R_A<\frac{3}{2}$. Starting from a more general list, Liang [4] improved this bound by stating that there is no such on-line algorithm in which $R_A<1.5364\cdots$.

By generalizing the result given in [4] we give a lower bound for lists in which $\max_{p_i \in L} l_i \leqq 1/r$ $(r=1,2,\cdots)$. Our result contains, as a special case, Liang's lower bound.

2. The parametric lower bound. Let L be a list and let r be the maximal integer on which $\max_{p_i \in L} l_i \leqq 1/r$. Let OPT (L) denote the minimal number of bins needed to pack the items, and $A(L)$ the number of bins that are used by the algorithm A to pack the list L. Let $R_A(k, r)$ denote the supremum of the quotients of $A(L)/\text{OPT}(L)$ for all lists with $\text{OPT}(L)=k$. Then the *asymptotic performance ratio* is $R_A(r)= \limsup_{k\to\infty} R_A(k, r)$.

During the proof of our theorem we use the following sequence: For a given integer r we define the elements of the sequence as follows:

$$t_1(r)=r+1, \quad t_{i+1}(r)=\prod_{j=1}^{i} t_j+1, \quad i \geqq 1.$$

We introduce further notation: let us suppose that $k \geqq r+1$, and let n be a suitable multiple of $t_{k+1}(r)-1$, and $0<\varepsilon \leqq 1/[(k+r-1)(t_{k+1}(r)-1)]$. Let us consider the list

* Received by the editors September 4, 1984, and in revised form July 17, 1985.

† Universität Augsburg, Institut für Mathematik, Memminger str. 6. D-8900 Augsburg, Federal Republic of Germany. On leave from Kalmár Laboratory of Cybernetics, József Attila University, Árpád tér 2 Szeged, H-6720, Hungary.

L_1 which contains nr pieces of elements of the same sizes $l_1 = 1/t_1(r) + \varepsilon$. Let L_j be lists which contain n pieces with sizes $l_j = 1/t_j(r) + \varepsilon$ $(2 \leqq j \leqq k)$. Finally we denote by $(L_k L_{k-1} \cdots L_j)$ the concatenated list in which the elements of L_k are followed by the elements of L_{k-1} etc.

LEMMA 1.

$$\mathrm{OPT}\,(L_k \cdots L_j) = \frac{n}{t_j(r) - 1}, \qquad j \geqq 2.$$

Proof. On the one hand

$$\sum_{i=j}^{k} \left(\frac{1}{t_i(r)} + \varepsilon \right) = \frac{1}{t_j(r) - 1} - \frac{1}{t_{k+1}(r) - 1} + (k + 1 - j)\varepsilon \leqq \frac{1}{t_j(r) - 1}.$$

So we can put the list into $n/(t_j(r) - 1)$ bins, therefore

$$\mathrm{OPT}\,(L_k \cdots L_j) \leqq \frac{n}{t_j(r) - 1}.$$

On the other hand, if we consider the elements of the list L_j only, it is true that we can pack at most $t_j(r) - 1$ pieces into one bin; thus

$$\mathrm{OPT}\,(L_k \cdots L_j) \geqq \frac{n}{t_j(r) - 1}.$$

The last two inequalities concerning the optimal packing give the desired result. □

LEMMA 2.

$$\mathrm{OPT}\,(L_k \cdots L_1) = n.$$

Proof. Consider a packing which places into one bin r pieces from the list L_1 and one piece from each of the other lists L_j $(2 \leqq j \leqq k)$. These pieces fit in one bin since the sum of them is:

$$\frac{r}{r+1} + \sum_{i=2}^{k} \frac{1}{t_i(r)} + (k + r - 1)\varepsilon \leqq 1.$$

Consequently,

$$\mathrm{OPT}\,(L_k \cdots L_1) \leqq n.$$

On the other hand, the sum of the pieces in this list is:

$$n \left(\frac{r}{r+1} + \sum_{i=2}^{k} \frac{1}{t_i(r)} + (k + r - 1)\varepsilon \right) \leqq n;$$

therefore

$$\mathrm{OPT}\,(L_k \cdots L_1) \geqq n.$$

The last two inequalities concerning the optimal packing give the statement of the lemma. □

Now we can formulate the following theorem:

THEOREM. *Let A be an optional on-line algorithm. Let us denote the quotient $n/[t_j(r) - 1]$ by α_j. Let $R_A^j(\alpha_j, r) = A(L_k \cdots L_j)/\mathrm{OPT}\,(L_k \cdots L_j)$, $(1 \leqq j \leqq k)$, $R_A^k(r) = \max_{1 \leqq j \leqq k} R_A^j(\alpha_j, r)$ and $\bar{R}_A(r) = \max_{k \to \infty} R_A^k(r)$. Then*

$$\bar{R}_A(r) = \left(1 + \sum_{j=2}^{\infty} \frac{j}{t_j(r) - 1} \right) \Big/ \left(1 + \sum_{j=2}^{\infty} \frac{1}{t_j(r) - 1} \right). \tag{1}$$

Proof. Since for all j $(2 \leqq j \leqq k)$ $R_A^j(\alpha_j, r) \leqq \bar{R}_A(r)$, so the maximum of the left-hand sides gives a lower bound for $\bar{R}_A(r)$.

First we examine the situation when we have packed all the elements preceding the list L_1, but no element from the list L_1 has been packed. Let us denote by $a_{(q,i_2,\cdots,i_k)}$ the number of those bins in which the algorithm A places q pieces from the list L_1 $(0 \leqq q \leqq r)$, i_2 pieces from L_2, i_3 pieces from L_3 etc. Then the following inequalities are true:

$$\sum_{q=0}^{r} i_j a_{(q,i_2,\cdots,i_k)} = n \qquad (2 \leqq j \leqq k), \tag{2}$$

$$\sum_{q=0}^{r} a_{(q,i_2,\cdots,i_k)} \geqq \frac{n}{r+1}. \tag{3}$$

Let us try to pack the elements of the list L_1 with the help of the algorithm A. We will denote by b_q the number of bins in which there is no element from the lists $L_k, \cdots, L_2$ and which contain exactly q pieces from the list L_1. Then

$$nr \leqq \sum_{q=1}^{r} q(a_{(q,i_2,\cdots,i_k)} + b_q), \tag{4}$$

and so

$$n \leqq \sum_{q=1}^{r} (a_{(q,i_2,\cdots,i_k)} + b_q). \tag{5}$$

Since $\bar{R}_A(r) \geqq A(L_k \cdots L_j)/\mathrm{OPT}\,(L_k \cdots L_j)$ $(j = 1, 2, \cdots, k)$, so in the case $j = 1$ we get

$$A(L_k \cdots L_1) = a_{(0,i_2,\cdots,i_k)} + \sum_{q=1}^{r} (a_{(q,i_2,\cdots,i_k)} + b_q) \leqq \bar{R}_A(r)n,$$

and using the inequality (5), we get

$$a_{(0,i_2,\cdots,i_k)} \leqq \bar{R}_A(r)n - \sum_{q=1}^{r} (a_{(q,i_2,\cdots,i_k)} + b_q) = \bar{R}_A(r)n - n. \tag{6}$$

Let us multiply both sides of the equalities (2) by $-j/t_j(r) - 1$:

$$-\sum_{q=0}^{r} \frac{j}{t_j(r)-1} i_j a_{(q,i_2,\cdots,i_k)} = -\frac{j}{t_j(r)-1} n. \tag{7}$$

The formula (7) is valid for all $2 \leqq j \leqq k$. If we use again the inequality concerning the lower bound, and we take into account the result of Lemma 1, we get

$$\sum_{q=0}^{r} a_{(q,i_2,\cdots,i_k)} \leqq \bar{R}_A(r) \frac{n}{t_j(r)-1}, \qquad 2 \leqq j \leqq k. \tag{8}$$

Adding all inequalities (6)–(8) we get

$$\begin{aligned} &\bar{R}_A(r)n - n + \bar{R}_A(r)n \sum_{j=2}^{k} \frac{1}{t_j(r)-1} - n \sum_{j=2}^{k} \frac{j}{t_j(r)-1} \\ &\geqq a_{(0,i_2,\cdots,i_k)} - \sum_{j=2}^{k} \sum_{q=0}^{r} \frac{j}{t_j(r)-1} i_j a_{(q,i_2,\cdots,i_k)} + \sum_{j=2}^{k} \sum_{q=0}^{r} a_{(q,i_2,\cdots,i_k)}, \end{aligned} \tag{9}$$

and we claim that the right-hand side of the inequality is nonnegative. If this is true, then rearranging the left-hand side, we get the statement of the theorem.

To see the nonnegativity, it is sufficient to show that the coefficient of each $a_{(q,i_2,\cdots,i_k)}$ $(0 \leqq q \leqq r)$ is nonnegative. We will distinguish two cases. Examining both of them, we can assume w.l.o.g. that $i_k > 0$, because if $i_k = 0$ we can replace k by the highest index k' for which $i_{k'} > 0$ and the rest of the proof remains unchanged.

Case 1. $q = 0$. Here we must prove that

$$\sum_{j=2}^{k} \frac{j}{t_j(r)-1} i_j \leqq k. \tag{10}$$

Our proof is based on two lemmas.

LEMMA 3. *Let us suppose that* $0 \leqq q \leqq r$. *If*

$$\sum_{j=2}^{k} \frac{i_j}{t_j(r)} < 1 - \frac{q}{r+1}$$

then

$$\sum_{j=2}^{k-1} \frac{i_j}{t_j(r)} + \frac{i_k}{t_k(r)-1} \leqq 1 - \frac{q}{r+1}.$$

Proof. If we suppose that in spite of the statement

$$\sum_{j=2}^{k-1} \frac{i_j}{t_j(r)} + \frac{i_k}{t_k(r)-1} > 1 - \frac{q}{r+1},$$

then, according to the rule making the $t_k(r)$ sequence, the left-hand side of this inequality is at least $1 - q/(r+1) + 1/(t_k(r)-1)$. So we must have

$$\begin{aligned}\sum_{j=2}^{k} \frac{i_j}{t_j(r)} &= \sum_{j=2}^{k-1} \frac{i_j}{t_j(r)} + \frac{i_k}{t_k(r)} + \frac{i_k}{t_k(r)-1} - \frac{i_k}{t_k(r)-1} \\ &\geqq 1 - \frac{q}{r+1} + \frac{1}{t_k(r)-1} - \left(\frac{i_k}{t_k(r)-1} - \frac{i_k}{t_k(r)}\right) \\ &\geqq 1 - \frac{q}{r+1} + \frac{1}{t_k(r)-1} - \frac{1}{t_k(r)-1} \\ &= 1 - \frac{q}{r+1},\end{aligned}$$

which is a contradiction. □

LEMMA 4. *If*

$$\sum_{j=2}^{k-1} \frac{i_j}{t_j(r)} + \frac{i_k}{t_k(r)-1} \leqq 1$$

then

$$\sum_{j=2}^{k} \frac{j}{t_j(r)-1} i_j \leqq k.$$

Proof. Since $j/(t_j(r)-1) \leqq k/t_j(r)$ if $2 \leqq j \leqq k-1$, thus

$$\sum_{j=2}^{k} \frac{j}{t_j(r)-1} i_j \leqq \sum_{j=2}^{k-1} \frac{k}{t_j(r)} i_j + \frac{k}{t_k(r)-1} i_k \leqq k. \qquad \square$$

Returning to the proof of nonnegativity in Case 1, we can see that the contents of the bins that belong to this case ($q=0$) are not greater than 1, i.e.

$$\sum_{j=2}^{k} \frac{i_j}{t_j(r)} + O(\varepsilon) \leqq 1,$$

from which

$$\sum_{j=2}^{k} \frac{i_j}{t_j(r)} < 1.$$

Using the above two lemmas, we get the inequality (10).

Case 2. $q>0$. We have to prove that

$$\sum_{j=2}^{k} \frac{j}{t_j(r)-1} i_j \leqq k-1. \tag{11}$$

Now we need a lemma again.

LEMMA 5. *If*

$$\sum_{j=2}^{k-1} \frac{i_j}{t_j(r)} + \frac{i_k}{t_k(r)-1} \leqq 1 - \frac{q}{r+1}$$

then

$$\sum_{j=2}^{k} \frac{j}{t_j(r)} i_j \leqq k-1.$$

Proof. Using up the inequality in Lemma 4, we get

$$\sum_{j=2}^{k} \frac{j}{t_j(r)} i_j \leqq \sum_{j=2}^{k-1} \frac{k}{t_j(r)} i_j + \frac{k}{t_k(r)-1} i_k \leqq k\left(1 - \frac{q}{r+1}\right) \leqq k-1.$$

The last inequality follows from the condition $k \geqq r+1$. □

Now we can easily prove the inequality (11). Since for the contents of the bins belonging to Case 2

$$\sum_{j=2}^{k} \frac{i_j}{t_j(r)} + O(\varepsilon) \leqq \frac{r-q+1}{r+1} = 1 - \frac{q}{r+1},$$

therefore

$$\sum_{j=2}^{k} \frac{i_j}{t_j(r)} < 1 - \frac{q}{r+1}.$$

TABLE 1
The values of $\bar{R}_A(r)$ for different k values.

$r-k$	2	3	4	5-∞
1	1.333	1.5	1.535	1.536
2		1.353	1.365	1.365
3			1.274	1.274
4				1.219

Using Lemmas 3 and 5, we get the inequality (11). So the proof of the theorem is completed. □

Finally, Table 1 shows some values of the lower bound for different k values, and Table 2 presents, over the lower bound value for $k=5$, the asymptotic performance ratio for some well-known algorithms.

TABLE 2
$R_A(r)$ values for different on-line algorithms.

	$r=1$	$r=2$	$r=3$	$r=4$
$\bar{R}_A(r)$	1.536	1.365	1.274	1.219
Worst Fit	2.0	2.0	1.5	1.33
Next Fit	2.0	2.0	1.5	1.33
Almost W.F.	2.0	2.0	1.5	1.33
First Fit	1.7	1.5	1.33	1.25
Harm. Fit	1.691	1.423	1.302	1.233
Rev. F.F.	1.66	N.A.	N.A.	N.A.

Acknowledgment. The author is grateful to a referee for suggesting Lemma 3 (a new and better version of a previous lemma).

REFERENCES

[1] B. S. BAKER AND E. G. COFFMAN [1981], *A tight asymptotic bound for next-fit-decreasing bin packing*, this Journal, 2, pp. 147-152.

[2] M. R. GAREY AND D. S. JOHNSON [1979], *Computer and Intractibility: A Guide to the Theory of* NP-*Completeness*, W. H. Freeman, San Francisco.

[3] D. S. JOHNSON, A. DEMERS, J. D. ULLMAN, M. R. GAREY AND R. L. GRAHAM [1974], *Worst-case performance bounds for simple one-dimensional packing algorithms*, SIAM J. Comput., 3, pp. 256-278.

[4] F. M. LIANG [1980], *Lower bound for on-line bin packing*, Inform. Proc. Lett., 10, pp. 76-79.

[5] A. C. C. YAO [1980], *New algorithms in bin packing*, J. ACM, 27, pp. 207-227.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

005

A STABLE METHOD FOR THE *LU* FACTORIZATION OF *M*-MATRICES*

ALAN A. AHAC† AND D. D. OLESKY‡

Abstract. We present an algorithm for the *LU* factorization of *M*-matrices based upon Gaussian elimination with a new pivoting strategy. At each step of the elimination, a column that is the most (column) diagonally dominant in the unreduced submatrix is exchanged into the pivotal column position through a symmetric permutation on the matrix. We demonstrate that this approach is well-suited to *M*-matrices, and can be implemented efficiently. The stability of the method is shown by providing a bound on the growth factor associated with the backward error analysis of the Gaussian elimination algorithm.

Key words. *M*-matrix, Gaussian elimination, pivoting, stability, *LU* factorization

AMS(MOS) subject classifications. 15A06, 15A23, 15A57, 65F05

1. Introduction. An $n\times n$ real matrix $A\equiv[a_{i,j}]$ is an *M-matrix* if $a_{i,j}\leqq 0$ for all $i\neq j$ and if $\operatorname{Re}\lambda\geqq 0$ for all $\lambda\in\sigma(A)$, the spectrum of A. It can be shown that an M-matrix A is nonsingular if and only if $\operatorname{Re}\lambda>0$ for all $\lambda\in\sigma(A)$; lengthy lists of necessary and sufficient conditions for A to be either a singular or a nonsingular M-matrix are given in Berman and Plemmons [1979].

An $n\times n$ matrix A is said to have an *LU factorization* if there exists a lower triangular matrix L and an upper triangular matrix U such that $A=LU$. Solving a system of linear equations

$$Ax=b \tag{1}$$

by Gaussian elimination involves the determination of an LU factorization of A (or PAQ, where P and Q denote permutation matrices). This paper is concerned with the determination of a permutation matrix P so that an LU factorization of the symmetric permutation PAP^T of an arbitrary (singular or nonsingular) M-matrix may be computed in a stable manner.

Letting $x=(x_1,x_2,\cdots,x_n)^T$ denote a vector of real numbers, we will use the following notation:

$x\geqq 0$ means $x_i\geqq 0$ for each i;
$x>0$ means $x_i\geqq 0$ and $x_i\neq 0$ for some i;
$x\gg 0$ means $x_i>0$ for each i.

This notation will also be used for matrix inequalities.

Several recent papers have considered the following problem: characterize those M-matrices A for which

(2) $A=LU$ with L a nonsingular lower triangular M-matrix and U an upper triangular M-matrix.

For nonsingular M-matrices, the factorization (2) follows from results in Fiedler and Ptak [1962], and this was extended to irreducible M-matrices by Kuo [1977] and to

* Received by the editors June 24, 1985, and in revised form July 17, 1985.

† Department of Computer Science, University of Victoria, Victoria, British Columbia, Canada V8W 2Y2. Current address, MacDonald Dettwiler and Associates, 3751 Shell Rd., Richmond, British Columbia, Canada V6X 2Z9.

‡ Department of Computer Science, University of Victoria, Victoria, British Columbia, Canada V8W 2Y2. The research of this author was supported by the Natural Sciences and Engineering Research Council of Canada under grant A-8214.

generalized diagonally dominant M-matrices by Funderlic and Plemmons [1981]. Note that A is *generalized diagonally dominant* if

$$(3) \qquad y^TA \geqq 0 \quad \text{for some vector } y \gg 0.$$

Graph-theoretic necessary and sufficient conditions for an M-matrix A to have a factorization (2) were given by Varga and Cai [1981].

The LU factorization of PAP^T, where A is an M-matrix and P a permutation matrix, is of particular interest since PAP^T is also an M-matrix, and thus retains the structure and properties of A. With regard to solving (1), the determination of an LU factorization of PAP^T may be realized by using Gaussian elimination with a pivoting strategy that employs simultaneous interchanges of identical rows and columns of the coefficient matrix at each step.

Although all (singular) M-matrices do not have an LU factorization, Kuo [1977] showed that for any M-matrix A there exists a permutation matrix P such that PAP^T has an LU factorization of the form (2). Varga and Cai [1981] proved that PAP^T has such an LU factorization for *all* permutation matrices P if and only if A is generalized diagonally dominant. Finally, generalized diagonal dominance of an M-matrix A has been shown to imply that $PBP^T = LU$ for all permutation matrices P and for all matrices B for which

$$|b_{j,j}| \geqq a_{j,j}, \qquad 1 \leqq j \leqq n,$$

and

$$|b_{i,j}| \leqq |a_{i,j}|, \qquad i \neq j$$

(see Funderlic, Neumann and Plemmons [1982]).

With regard to the above results, the algorithm which we describe determines a permutation matrix P and matrices L and U as in (2) such that $PAP^T = LU$, where A is an arbitrary M-matrix; thus it is a realization of Kuo's existence theorem. Furthermore, the algorithm is shown to be stable and efficient with respect to the number of arithmetic operations required for its implementation.

2. Stability of Gaussian elimination. Let A be an $n \times n$ matrix. We will denote the reduced matrix which is determined after k steps $(1 \leqq k \leqq n-1)$ of the forward elimination of Gaussian elimination (which we abbreviate to GE) by

$$(4) \qquad A_k = \left[\begin{array}{cccccccc}
a_{1,1} & a_{1,2} & \cdot & a_{1,k} & a_{1,k+1} & \cdot & \cdot & a_{1,n} \\
 & a_{2,2}^{(1)} & \cdot & a_{2,k}^{(1)} & a_{2,k+1}^{(1)} & \cdot & \cdot & a_{2,n}^{(1)} \\
 & & \cdot & \cdot & \cdot & & & \cdot \\
 & & & a_{k,k}^{(k-1)} & a_{k,k+1}^{(k-1)} & \cdot & \cdot & a_{k,n}^{(k-1)} \\
 & 0 & & & a_{k+1,k+1}^{(k)} & \cdot & \cdot & a_{k+1,n}^{(k)} \\
 & & & & a_{k+2,k+1}^{(k)} & \cdot & & a_{k+2,n}^{(k)} \\
 & & & & \cdot & & \cdot & \cdot \\
 & & & & a_{n,k+1}^{(k)} & \cdot & \cdot & a_{n,n}^{(k)}
\end{array}\right].$$

Thus A_{n-1} is upper triangular. For convenience we denote A by A_0, and let $a_{i,j} \equiv a_{i,j}^{(0)}$. When GE with floating point arithmetic is applied to A, triangular factors L and A_{n-1} are determined such that

$$(5) \qquad LA_{n-1} = A + E.$$

This backward error analysis is due to Wilkinson [1961] and bounds on the magnitude of elements of E are given, for example, by Reid [1971], and involve

$$a=\max_{i,j,k}|a_{i,j}^{(k)}|,\qquad 0\leqq k\leqq n-1,$$

the largest element in magnitude in any of the reduced matrices A_k. It is common, and often more meaningful, to consider the *growth factor* γ defined as

$$\gamma=\frac{\max_{i,j,k}|a_{i,j}^{(k)}|}{\max_{i,j}|a_{i,j}|},\qquad 0\leqq k\leqq n-1,$$

as a measure of the size of the perturbation E with respect to the magnitude of the elements of A. Clearly $\gamma\geqq 1$.

We will say that an algorithm using floating point arithmetic is stable if the algorithm yields a solution that is "near" the exact solution of a slightly perturbed problem (see e.g. Stewart [1973]). Thus, if the elements of the matrix E of (5) can be shown to be "small," then the GE algorithm can be considered stable. This essentially requires controlling the size of γ.

For GE with partial pivoting, it is well known that the best possible bound is $\gamma\leqq 2^{n-1}$, while for complete pivoting (see e.g. Stewart [1973])

$$\gamma<(n\cdot 2^1 3^{1/2} 4^{1/3}\cdots n^{1/(n-1)})^{1/2}.$$

For certain classes of matrices, the growth factor may be bounded independently of n. For example, if A is either column diagonally dominant or tridiagonal, then $\gamma\leqq 2$ (using GE with no pivoting or with partial pivoting, respectively) while if A is positive definite, then $\gamma\leqq 1$ (without pivoting).

3. *LU* factorization of an *M*-matrix.

We will write an M-matrix A as

$$(6)\qquad A\equiv A_0=\begin{bmatrix} a_{1,1} & -a_{1,2} & -a_{1,3} & \cdot & \cdot & \cdot \\ -a_{2,1} & a_{2,2} & -a_{2,3} & \cdot & \cdot & \cdot \\ -a_{3,1} & -a_{3,2} & a_{3,3} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & & \cdot \end{bmatrix},$$

where $a_{i,j}\geqq 0$ for all i,j. An M-matrix that has an LU factorization is invariant with respect to the application of GE without pivoting (see e.g. Fan [1960]). Thus the reduced matrix (4) may be written as

$$(7)\qquad A_k=\begin{bmatrix} a_{1,1} & -a_{1,2} & \cdot & -a_{1,k} & -a_{1,k+1} & \cdot & \cdot & -a_{1,n} \\ & a_{2,2}^{(1)} & \cdot & -a_{2,k}^{(1)} & -a_{2,k+1}^{(1)} & \cdot & \cdot & -a_{2,n}^{(1)} \\ & & \cdot & \cdot & \cdot & & & \cdot \\ & & & a_{k,k}^{(k-1)} & -a_{k,k+1}^{(k-1)} & \cdot & \cdot & -a_{k,n}^{(k-1)} \\ & & & & a_{k+1,k+1}^{(k)} & \cdot & \cdot & -a_{k+1,n}^{(k)} \\ & 0 & & & -a_{k+2,k+1}^{(k)} & \cdot & & -a_{k+2,n}^{(k)} \\ & & & & \cdot & & \cdot & \cdot \\ & & & & -a_{n,k+1}^{(k)} & \cdot & \cdot & a_{n,n}^{(k)} \end{bmatrix}=\begin{bmatrix} \hat{U}_k & \hat{C}_k \\ 0 & \hat{A}_k \end{bmatrix}$$

where $\hat{U}_k$ is a $k\times k$ upper triangular M-matrix, $\hat{A}_k$ is an $(n-k)\times(n-k)$ M-matrix and $\hat{C}_k$ is a $k\times(n-k)$ matrix of nonpositive elements.

Funderlic and Plemmons [1981] note that if A satisfies (3), then DA is column diagonally dominant, where $D=\operatorname{diag}(y_1,y_2,\cdots,y_n)$, so that GE without pivoting applied to DA is stable (with $\gamma = 1$). In Funderlic, Neumann and Plemmons [1982], it is shown that GE without pivoting may be applied to A with

$$\gamma \leqq \frac{\max_i y_i}{\min_i y_i}.$$

However, in general y is unknown and the above bound on γ may be large. For example, the LU factorization

$$\begin{bmatrix} 2 & 0 & -100 \\ -100 & 100 & -1 \\ 0 & -1 & 100 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -50 & 1 & 0 \\ 0 & -0.01 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & -100 \\ 0 & 100 & -5001 \\ 0 & 0 & 49.99 \end{bmatrix}$$

has $\gamma = 50.01$.

The use of partial or complete pivoting to factor an M-matrix is unattractive since the reduced matrices A_k would not, in general, be M-matrices, and the desirable M-matrix properties and sign pattern would be lost. However, if A is an M-matrix, then so is a symmetric permutation PAP^T, and by Kuo [1977] we know that for *every* M-matrix A there exists a permutation matrix P such that PAP^T has an LU factorization (2). These comments motivate the pivoting strategy of the next section.

4. Column diagonal dominant pivoting. The following theorem states that there is always at least one column of an M-matrix that satisfies the column diagonal dominance condition. But before proceeding we need a definition.

For a general $n \times n$ matrix A, let $\tilde{A}$ denote the *reduced normal form* of A. That is,

$$\tilde{A} = PAP^T = \begin{bmatrix} \tilde{A}_{1,1} & \tilde{A}_{1,2} & \cdots & \tilde{A}_{1,k} \\ & \tilde{A}_{2,2} & \cdots & \tilde{A}_{2,k} \\ & & \ddots & \vdots \\ & 0 & & \tilde{A}_{k,k} \end{bmatrix},$$

where P is a permutation matrix and $\tilde{A}_{j,j}$ is either irreducible or a zero matrix of order one, $1 \leqq j \leqq k$.

THEOREM 1. *Given an $n \times n$ M-matrix A, there exists at least one subscript j such that*

$$a_{j,j} \geqq \sum_{\substack{i=1 \\ i \neq j}}^{n} a_{i,j}. \tag{8}$$

Proof. Assume without loss of generality that A is in reduced normal form. Then if $\tilde{A}_{1,1}$ is a zero matrix of order one, trivially

$$0 = a_{1,1} \geqq \sum_{i=2}^{n} a_{i,1} = 0$$

and the condition is satisfied. If $\tilde{A}_{1,1} \neq 0$, it is sufficient to show that the condition is true for an irreducible M-matrix A.

Any M-matrix may be written as $A = sI - B$ with $B \geqq 0$ and $s \geqq \rho(B)$, the spectral radius of B. We assume A is irreducible, so B is irreducible as well. Let j be a column of B such that

$$\sum_{i=1}^{n} b_{i,j} = \min_{1 \leqq l \leqq n} \left(\sum_{i=1}^{n} b_{i,l} \right).$$

Then by a well-known result of Perron-Frobenius theory (see Varga [1962, Lemma

2.5, p. 31])

$$\rho(B) \geqq \sum_{i=1}^{n} b_{i,j}$$

and $s \geqq \rho(B)$ implies

$$s - b_{j,j} \geqq \sum_{\substack{i=1 \\ i \neq j}}^{n} b_{i,j}$$

which is (8). □

Before proceeding, we note two special cases of the above result.

THEOREM 2. *Given a $n \times n$ nonsingular M-matrix A, there exists at least one subscript j such that*

$$a_{j,j} > \sum_{\substack{i=1 \\ i \neq j}}^{n} a_{i,j}.$$

Proof. Since A is nonsingular, the reduced normal form of A cannot have $\tilde{A}_{1,1}$ as a zero matrix of order one, so we need only consider the irreducible (nonsingular) case.

The result now follows from the proof of Theorem 1 when we note that $s > \rho(B)$ in the nonsingular case. □

THEOREM 3. *Let A be an $n \times n$ singular, irreducible M-matrix. There exists a column j such that*

$$a_{j,j} > \sum_{\substack{i=1 \\ i \neq j}}^{n} a_{i,j} \tag{9}$$

if and only if there exists a column r such that

$$a_{r,r} < \sum_{\substack{i=1 \\ i \neq r}}^{n} a_{i,r}. \tag{10}$$

Proof. Assume firstly that (9) is satisfied. If there is no column r as in (10), then

$$a_{r,r} \geqq \sum_{\substack{i=1 \\ i \neq r}}^{n} a_{i,r}$$

for all $r \neq j$. But then $(1, \cdots, 1)A \geqq 0$, and since A is *almost monotone* (see Berman and Plemmons [1979, p. 156]), this implies $(1, \cdots, 1)A = 0$, which contradicts (9).

To prove the converse, assume that (10) is satisfied and that there is no j for which (9) holds. A contradiction is obtained as $(-1, \cdots, -1)A = 0$. □

We now define a pivoting strategy based on the existence of a diagonally dominant column. Assume $A = A_0$ is any M-matrix (6), let A_k denote the reduced M-matrix (7), and assume that column $j_k (k < j_k \leqq n)$ has the *maximal* column sum in $\hat{A}_k$. (By Theorem 1, this sum is nonnegative.) Then *column diagonal dominant pivoting* (*cdd-pivoting*) is the process of exchanging the $(k+1)$th and j_kth columns and rows of A_k prior to the $(k+1)$th step of $GE (0 \leqq k \leqq n-2)$. The interchanging is equivalent to forming

$$P_{k+1,j_k} A_k P_{k+1,j_k}^T$$

where P_{k+1,j_k} is an elementary permutation matrix.

Note that the application of *GE* with cdd-pivoting implies that the *sum* of the absolute values of the multipliers at each step is $\leqq 1$. For *GE* with partial or complete pivoting, we know only that the magnitude of *each* multiplier is $\leqq 1$.

In general, *GE* (without pivoting) will fail if the elimination process results in $a^{(k)}_{k+1,k+1}=0$ and $a^{(k)}_{l,k+1}\neq 0$ for some $l\in\{k+2,\cdots,n\}$. Using cdd-pivoting on M-matrices ensures that the pivotal column sum is nonnegative so if a zero pivot element is encountered, then $a^{(k)}_{l,k+1}=0$ for all $l\in\{k+2,\cdots,n\}$. No computation is needed at that step, and we can continue on to the $(k+1)$th step. Thus, *GE* with cdd-pivoting will not fail for any M-matrix, and the resultant LU factorization will be an LU factorization of PAP^T for some permutation matrix P.

The following theorem relates the column sums of the unreduced submatrix $\hat{A}_1$ to the column sums of $A=A_0$. This result provides an easy method for determining the column sums of the unreduced submatrix $\hat{A}_k$ at the kth step in the elimination based on information computed at the previous step.

THEOREM 4. *Let A be an $n\times n$ M-matrix, and define*

$$s_j=a_{j,j}-\sum_{\substack{i=1\\i\neq j}}^{n}a_{i,j},\qquad 1\leqq j\leqq n. \tag{11}$$

Let A_1 (cf. (7)) be the result of one step of GE applied to A and define

$$s_j^{(1)}=a_{j,j}^{(1)}-\sum_{\substack{i=2\\i\neq j}}^{n}a_{i,j}^{(1)},\qquad 2\leqq j\leqq n. \tag{12}$$

Then

$$s_j^{(1)}=s_j+s_1\left(\frac{a_{1,j}}{a_{1,1}}\right),\qquad 2\leqq j\leqq n.$$

Proof. Firstly we note from (11)

$$\sum_{\substack{i=1\\i\neq j}}^{n}\frac{a_{i,j}}{a_{j,j}}=1-\frac{s_j}{a_{j,j}},\qquad 1\leqq j\leqq n.$$

From (12),

$$\begin{aligned}s_j^{(1)}&=a_{j,j}-\frac{a_{j,1}}{a_{1,1}}a_{1,j}-\sum_{\substack{i=2\\i\neq j}}^{n}\left(a_{i,j}+\frac{a_{i,1}}{a_{1,1}}a_{1,j}\right)\\&=a_{j,j}-\sum_{\substack{i=2\\i\neq j}}^{n}a_{i,j}-a_{1,j}\sum_{i=2}^{n}\frac{a_{i,1}}{a_{1,1}}\\&=a_{j,j}-\sum_{\substack{i=2\\i\neq j}}^{n}a_{i,j}-a_{1,j}\left(1-\frac{s_1}{a_{1,1}}\right)\\&=a_{j,j}-\sum_{\substack{i=1\\i\neq j}}^{n}a_{i,j}+s_1\frac{a_{1,j}}{a_{1,1}}\\&=s_j+s_1\frac{a_{1,j}}{a_{1,1}},\qquad 2\leqq j\leqq n.\end{aligned}$$

□

5. Stability of Gaussian elimination with cdd-pivoting. To examine the stability of our algorithm, we need to find a bound on the growth factor γ for the method. The interchanges required for the pivoting strategy complicate the expressions for the A_k, making analysis of the algorithm difficult. Fortunately, the following lemma shows that using *GE* with cdd-pivoting is equivalent to doing the interchanges first and then applying *GE* without any pivoting.

LEMMA 1. *Let $A' = A'_0$ be an $n \times n$ M-matrix and let $1 \leqq k \leqq n-1$. Let M'_i, $1 \leqq i \leqq k$, denote elementary lower triangular matrices and P_i, $1 \leqq i \leqq k$, denote elementary permutation matrices such that*

$$A'_i = M'_i \, P_i A'_{i-1} P_i^T, \qquad 1 \leqq i \leqq k,$$

denote the reduced matrices obtained by applying k steps of GE with cdd-pivoting to A'. Let $A_0 = P_k P_{k-1} \cdots P_1 A'_0 P_1^T \cdots P_{k-1}^T P_k^T$. Then GE without pivoting may be applied to A_0 (and if some pivot $a_{i,i}^{(i-1)} = 0$, then $a_{j,i}^{(i-1)} = 0$ for $i+1 \leqq j \leqq n$). If M_i, $1 \leqq i \leqq k$, denote the elementary lower triangular matrices such that

$$A_i = M_i A_{i-1}, \qquad 1 \leqq i \leqq k,$$

are the reduced matrices obtained by applying k steps of GE without pivoting to A_0, then

$$A_k = A'_k.$$

Proof. See Stewart [1973, Thm. 2.9, p. 125]. □

Using Lemma 1, we obtain the following result which gives an upper bound, dependent upon values contained in $A = A_0, \cdots, A_{k-2}$, for the elements of the intermediate matrix A_k, $1 \leqq k \leqq n-1$.

LEMMA 2. *Let A' be an $n \times n$ nonsingular M-matrix and let $1 \leqq k \leqq n-2$. Let $A = A_0$ and A_i, $1 \leqq i \leqq k$, be defined as in Lemma 1. Then for $1 \leqq t \leqq k-1$,*

$$\begin{aligned} a_{i,j}^{(k)} < a_{i,j} &+ \sum_{s=k-t+1}^{k} a_{s,j} + a_{1,j}\left(\frac{a_{i,1}}{a_{1,1}} + \sum_{s=k-t+1}^{k} \frac{a_{s,1}}{a_{1,1}}\right) \\ &+ \sum_{l=1}^{k-t-1} a_{l+1,j}^{(l)} \left(\frac{a_{i,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}} + \sum_{r=k-t+1}^{k} \frac{a_{r,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}}\right) \end{aligned} \tag{13}$$

where $k+1 \leqq i \leqq n$, $k+1 \leqq j \leqq n$ and $i \neq j$.

Proof. We first note that all elements which appear in denominators in (13) are nonzero since the reduced matrices A_k are also nonsingular M-matrices.

The off-diagonal elements of A are defined by

$$-a_{i,j}^{(k)} = -a_{i,j}^{(k-1)} - \frac{a_{i,k}^{(k-1)}}{a_{k,k}^{(k-1)}} a_{k,j}^{(k-1)},$$

from which it follows that

$$a_{i,j}^{(k)} = a_{i,j} + \frac{a_{i,1}}{a_{1,1}} a_{1,j} + \sum_{l=1}^{k-1} a_{l+1,j}^{(l)} \frac{a_{i,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}}. \tag{14}$$

The lemma is now proved by induction on t.

From (14) we obtain

$$\begin{aligned} a_{i,j}^{(k)} &= a_{i,j} + \frac{a_{i,1}}{a_{1,1}} a_{1,j} + \frac{a_{i,k}^{(k-1)}}{a_{k,k}^{(k-1)}} a_{k,j}^{(k-1)} + \sum_{l=1}^{k-2} a_{l+1,j}^{(l)} \frac{a_{i,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}} \\ &< a_{i,j} + \frac{a_{i,1}}{a_{1,1}} a_{1,j} + a_{k,j}^{(k-1)} + \sum_{l=1}^{k-2} a_{l+1,j}^{(l)} \frac{a_{i,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}} \end{aligned}$$

since cdd-pivoting guarantees that $a_{i,k}^{(k-1)}/a_{k,k}^{(k-1)}<1$. On expanding $a_{k,j}^{(k-1)}$ using (14),

$$\begin{aligned}
a_{i,j}^{(k)} &< a_{i,j}+\frac{a_{i,1}}{a_{1,1}}a_{1,j}+a_{k,j}+\frac{a_{k,1}}{a_{1,1}}a_{1,j}\\
&\quad+\sum_{l=1}^{k-2}a_{l+1,j}^{(l)}\frac{a_{k,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}}+\sum_{l=1}^{k-2}a_{l+1,j}^{(l)}\frac{a_{i,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}}\\
&=a_{i,j}+a_{k,j}+a_{1,j}\left(\frac{a_{i,1}}{a_{1,1}}+\frac{a_{k,1}}{a_{1,1}}\right)\\
&\quad+\sum_{l=1}^{k-2}a_{l+1,j}^{(l)}\left(\frac{a_{i,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}}+\frac{a_{k,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}}\right).
\end{aligned}$$

Thus, the condition is true for $t=1$.

Now, on assuming that (13) holds for $t=m-1<k-1$, we have

$$\begin{aligned}
a_{i,j}^{(k)} &< a_{i,j}+\sum_{s=k-m+2}^{k}a_{s,j}+a_{1,j}\left(\frac{a_{i,1}}{a_{1,1}}+\sum_{s=k-m+2}^{k}\frac{a_{s,1}}{a_{1,1}}\right)\\
&\quad+\sum_{l=1}^{k-m}a_{l+1,j}^{(l)}\left(\frac{a_{i,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}}+\sum_{r=k-m+2}^{k}\frac{a_{r,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}}\right)\\
&=a_{i,j}+\sum_{s=k-m+2}^{k}a_{s,j}+a_{1,j}\left(\frac{a_{i,1}}{a_{1,1}}+\sum_{s=k-m+2}^{k}\frac{a_{s,1}}{a_{1,1}}\right)\\
&\quad+a_{k-m+1,j}^{(k-m)}\left(\frac{a_{i,k-m+1}^{(k-m)}}{a_{k-m+1,k-m+1}^{(k-m)}}+\sum_{r=k-m+2}^{k}\frac{a_{r,k-m+1}^{(k-m)}}{a_{k-m+1,k-m+1}^{(k-m)}}\right)\\
&\quad+\sum_{l=1}^{k-m-1}a_{l+1,j}^{(l)}\left(\frac{a_{i,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}}+\sum_{r=k-m+2}^{k}\frac{a_{r,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}}\right).
\end{aligned}$$

Cdd-pivoting implies that

$$\left(\frac{a_{i,k-m+1}^{(k-m)}}{a_{k-m+1,k-m+1}^{(k-m)}}+\sum_{r=k-m+2}^{k}\frac{a_{r,k-m+1}^{(k-m)}}{a_{k-m+1,k-m+1}^{(k-m)}}\right)<1,$$

and expansion of $a_{k-m+1,j}^{(k-m)}$ using (14) yields

$$\begin{aligned}
a_{i,j}^{(k)} &< a_{i,j}+\sum_{s=k-m+2}^{k}a_{s,j}+a_{1,j}\left(\frac{a_{i,1}}{a_{1,1}}+\sum_{s=k-m+2}^{k}\frac{a_{s,1}}{a_{1,1}}\right)+a_{k-m+1,j}\\
&\quad+\frac{a_{k-m+1,1}}{a_{1,1}}a_{1,j}+\sum_{l=1}^{k-m-1}a_{l+1,j}^{(l)}\frac{a_{k-m+1,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}}\\
&\quad+\sum_{l=1}^{k-m-1}a_{l+1,j}^{(l)}\left(\frac{a_{i,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}}+\sum_{r=k-m+2}^{k}\frac{a_{r,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}}\right)\\
&=a_{i,j}+\sum_{s=k-m+1}^{k}a_{s,j}+a_{1,j}\left(\frac{a_{i,1}}{a_{1,1}}+\sum_{s=k-m+1}^{k}\frac{a_{s,1}}{a_{1,1}}\right)\\
&\quad+\sum_{l=1}^{k-m-1}a_{l+1,j}^{(l)}\left(\frac{a_{i,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}}+\sum_{r=k-m+1}^{k}\frac{a_{r,l+1}^{(l)}}{a_{l+1,l+1}^{(l)}}\right)
\end{aligned}$$

which is (13) with $t=m$, completing the proof. □

We can now state our result regarding the stability of *GE* with cdd-pivoting on *M*-matrices.

THEOREM 5. *Let A be an $n\times n$ nonsingular M-matrix. Then the growth factor resulting from the application of GE with cdd-pivoting to A is bounded by*

$$\gamma<n-1. \tag{15}$$

Proof. Since growth in the elements of the reduced matrices produced by *GE* only occurs in off-diagonal positions, we need only consider bounding the growth of off-diagonal elements of A_k, $1\leqq k\leqq n-2$.

Letting $t=k-1$ in (13) we have

$$a_{i,j}^{(k)}<a_{i,j}+\sum_{s=2}^{k}a_{s,j}+a_{1,j}\left(\frac{a_{i,1}}{a_{1,1}}+\sum_{s=2}^{k}\frac{a_{s,1}}{a_{1,1}}\right),$$

where $k+1\leqq i\leqq n$, $k+1\leqq j\leqq n$ and $i\neq j$. Since

$$\left(\frac{a_{i,1}}{a_{1,1}}+\sum_{s=2}^{k}\frac{a_{s,1}}{a_{1,1}}\right)<1$$

when using cdd-pivoting,

$$a_{i,j}^{(k)}<a_{i,j}+\sum_{s=1}^{k}a_{s,j}<(k+1)\max_{s\in\{1,\cdots,k,i\}}a_{s,j}.$$

The result now follows on letting $k=n-2$, since no growth in off-diagonal elements can occur when $k=n-1$. □

The matrix

$$A_0=\begin{bmatrix} 1 & 0 & \cdot & \cdot & 0 & 0 & -1\\ -1+\varepsilon & 1 & \cdot & & 0 & 0 & -1\\ & -1+\varepsilon & \cdot & \cdot & \cdot & \cdot & \cdot\\ & & \cdot & \cdot & 0 & 0 & -1\\ & & & \cdot & 1 & -1 & -1\\ & 0 & & & -1+\varepsilon & 1 & -1\\ & & & & 0 & -\hat{\varepsilon} & 1 \end{bmatrix} \tag{16}$$

is an $n\times n$ nonsingular irreducible M-matrix for $0<\varepsilon\leqq 1$ and for sufficiently small $\hat{\varepsilon}>0$. Applying *GE* with cdd-pivoting results in

$$A_0=\begin{bmatrix} 1 & & & & & & \\ -1+\varepsilon & 1 & & & 0 & & \\ 0 & -1+\varepsilon & \cdot & & & & \\ 0 & 0 & & \cdot & & & \\ \cdot & \cdot & & & 1 & & \\ \cdot & \cdot & & & -1+\varepsilon & 1 & \\ 0 & 0 & \cdot & \cdot & 0 & -\hat{\varepsilon}/\varepsilon & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & \cdot & \cdot & 0 & 0 & -1\\ & 1 & \cdot & & \cdot & \cdot & -2+\varepsilon\\ & & \cdot & \cdot & \cdot & \cdot & -3+\delta_3(\varepsilon)\\ & & & \cdot & 0 & 0 & -4+\delta_4(\varepsilon)\\ & & & & 1 & -1 & \cdot\\ & 0 & & & & \varepsilon & -(n-1)+\delta_{n-1}(\varepsilon)\\ & & & & & & 1-(\hat{\varepsilon}/\varepsilon)[n-1-\delta_{n-1}(\varepsilon)] \end{bmatrix}$$

$$\text{where } \delta_i(\varepsilon)=\frac{i(i-1)}{2}\varepsilon+O(\varepsilon^2).$$

Thus for sufficiently small $\varepsilon>0$, we can obtain growth in the $(n-1,n)$-position of U approaching $n-1$ times the maximum element in magnitude in A_0.

The strict inequality in (15) for nonsingular M-matrices is also applicable to singular, irreducible M-matrices. By Theorem 3, if A is a singular irreducible M-matrix, it has all column sums equal to zero, or it has at least one negative column sum and at least one positive column sum. For maximal growth to occur, the latter must be true. But this implies that there exists a pivotal column with positive column sum, indicating that the sum of the multipliers at that step will be strictly less than one. This ensures that the growth factor γ cannot equal $n-1$. Note that by choosing $\hat{\varepsilon}$ in (16) so that the (n, n)-element of U is zero, A_0 is a singular, irreducible M-matrix and the upper bound of $n-1$ on γ is seen to be tight for this case.

We observe that the 5×5 singular reducible M-matrix

$$A_0=\begin{bmatrix} 1 & 0 & 0 & -1 & -1 \\ -1 & 1 & 0 & 0 & -1 \\ 0 & -1 & 1 & 0 & -1 \\ 0 & 0 & -1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

can be decomposed using GE with cdd-pivoting to return the lower and upper triangular factors

$$\begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ 0 & -1 & 1 & & \\ 0 & 0 & -1 & 1 & \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 & -1 & -1 \\ & 1 & 0 & -1 & -2 \\ & & 1 & -1 & -3 \\ & & & 0 & -4 \\ & & & & 1 \end{bmatrix}.$$

The growth factor γ is 4 in this example, so our bound of Theorem 5 can be attained for singular reducible M-matrices. This is consistent with our expectations considering that Theorem 1 shows that the sum of the multipliers at every step of GE can be equal to one. Note that the above example generalizes to the $n\times n$ case.

6. Concluding remarks. We have presented a new pivoting strategy for computing the LU factorization (with L nonsingular) of a symmetric permutation PAP^T of an arbitrary M-matrix. The use of cdd-pivoting assures stability since the growth factor γ associated with the backward error analysis of GE is bounded by $n-1$. Previous algorithms either assume that A is column diagonally dominant or else rely on knowledge of a scaling vector y (see (3)) to insure stability; bounds on γ involving y (e.g. Funderlic, Neumann and Plemmons [1982]) may be very pessimistic.

The floating-point operation counts for the LU factorization of an $n\times n$ M-matrix using cdd-pivoting are

$$\tfrac{1}{3}n^3+\tfrac{1}{2}n^2+\tfrac{1}{6}n-3 \text{ multiplications/divisions}$$

and

$$\tfrac{1}{3}n^3+n^2-\tfrac{4}{3}n-1 \text{ additions/subtractions,}$$

which are, respectively, approximately $n^2/2$ and $3n^2/2$ more operations than are required for GE with either partial or no pivoting.

Finally, we note that for large, sparse matrices our algorithm should be modified to attempt to reduce the matrix fill-in. Since the bound $\gamma\leqq n-1$ depends only on the sum of the absolute values of the multipliers being less than or equal to one, the pivotal column can be any column of the unreduced submatrix $\hat{A}_k$ having a positive column sum (not necessarily the maximal column sum). Thus, for example, the sparsity criterion attributed to Markowitz (see Duff and Reid [1979]) may be used: at each step of the elimination, the jth column of $\hat{A}_k$ (see (7)) is chosen as the pivotal column, where

$$a_{j,j}^{(k)} \geqq \sum_{\substack{i=k+1 \\ i \neq j}}^{n} a_{i,j}^{(k)}$$

and the number of nonzero terms $a_{i,j}^{(k)} a_{j,i}^{(k)}$, $k+1 \leqq i \leqq n$, is a minimum. Note, in addition, that we can pivot solely for sparsity (without regard for stability) if at any stage of the elimination all of the column sums of $\hat{A}_k$ are nonnegative (since it then follows from Theorem 4 that all column sums of $\hat{A}_j$ must be nonnegative for all $j > k$).

REFERENCES

A. Berman and R. J. Plemmons [1979], *Nonnegative Matrices in the Mathematical Sciences*, Series on Computer Science and Applied Mathematics, Academic Press, New York.

I. S. Duff and J. K. Reid [1979], *Some design features of a sparse matrix code*, ACM Trans. Math. Software, 5, pp. 18–35.

K. Fan [1960], *Note on M-matrices*, Quart. J. Math. Oxford Ser. (2), 11, pp. 43–49.

M. Fiedler and V. Ptak [1962], *On matrices with nonpositive off-diagonal elements and positive principal minors*, Czech. Math. J., 12, pp. 382–400.

R. E. Funderlic, M. Neumann and R. J. Plemmons [1982], *LU decompositions of generalized diagonally dominant matrices*, Numer. Math., 40, pp. 57–69.

R. E. Funderlic and R. J. Plemmons [1981], *LU decomposition of M-matrices by elimination without pivoting*, Linear Algebra Appl., 41, pp. 99–110.

I. Kuo [1977], *A note on factorizations of singular M-matrices*, Linear Algebra Appl., 16, pp. 217–220.

J. K. Reid [1971], *A note on the stability of Gaussian elimination*, J. Inst. Maths Appl., 8, pp. 374–375.

G. W. Stewart [1973], *Introduction to Matrix Computations*, Series on Computer Science and Applied Mathematics, Academic Press, New York.

R. S. Varga [1962], *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ.

R. S. Varga and D.-Y. Cai [1981], *On the LU factorization of M-matrices*, Numer. Math., 38, pp. 179–192.

J. Wilkinson [1961], *Error analysis of direct methods of matrix inversion*, J. ACM, 8, pp. 281–330.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

006

EFFICIENT ALGORITHMS FOR OPTIMIZATION AND SELECTION ON SERIES-PARALLEL GRAPHS*

R. HASSIN† AND A. TAMIR†

Abstract. It is well known that a series-parallel multigraph G can be constructed recursively from its edges. This construction is represented by a binary decomposition tree. This is a rooted binary tree T in which each vertex q corresponds to some series-parallel submultigraph of G, denoted by $G(q)$, obtained as follows. Each leaf (tip) of T represents a distinct edge of G. If q is not a leaf then it is either of a series or parallel type. If q_1 and q_2 are the two sons of q on T, $G(q)$ is the submultigraph obtained from $G(q_1)$ and $G(q_2)$ by the respective series or parallel composition.

In this paper we use this tree to develop efficient algorithms for several optimization and selection problems defined on graphs with no K_4 homeomorph. In particular, we provide a linear time algorithm to find the shortest simple paths from a given vertex to all other vertices. We also construct an $O(n^4)$ algorithm to solve the uncapacitated plant location problem, where n is the number of vertices.

We prove that graphs with no K_4 homeomorph contain a 2-separator, and then use it to show that the kth longest path in the set of all intervertex distances can be selected in $O(n \log^3 n)$ time.

Key words. series-parallel graphs, graph decomposition, selection algorithms, uncapacitated plant location problem

AMS(MOS) subject classification. 90C35

1. Introduction. It is well known that various optimization problems that are NP-hard on general graphs can be solved efficiently on trees, by polynomial algorithms. Some of these algorithms (e.g. [KH], [KT], [MTZC]) are of the dynamic programming nature. They represent the given tree as a rooted tree, and then recursively solve subproblems associated with a sequence of subtrees, (while starting with the tips of the rooted tree). In other words, these algorithms rely heavily upon the existence of an efficient construction that recursively generates larger components from previous ones, and terminate with the given tree, T. The key property of this construction is that each component T_i, generated in the process, is associated with some distinguished vertex, say v_i, such that, every simple path connecting a vertex in T_i with a vertex in $T - T_i$ must contain v_i.

Another property of trees which is useful in designing efficient divide-and-conquer algorithms, is the existence of a 1-separator (e.g. [GOL], [KH]). A 1-separator of an n-vertex tree $T = (V, E)$ is a vertex v for which there exists a partition of $V - \{v\}$ consisting of V_1 and V_2 with,

(1) $V_1 \cap V_2 = \emptyset$,

(2) $|V_i| \leqq (2/3)n$, $i = 1, 2$, and

(3) no edge in E connects a vertex in V_1 with a vertex in V_2.

Another class of graphs for which an efficient construction (or decomposition) scheme is available is the class of series-parallel graphs [VTL]. Such a construction is represented by a binary decomposition tree that can be found in linear time [VTL]. Here, each component G_i of the underlying graph G, generated in the process is associated with two vertices, say $\{u_i, v_i\}$ such that every simple path connecting a vertex in G_i with a vertex in $G - G_i$, must contain either u_i or v_i.

In this paper we focus on graphs, not necessarily biconnected, that contain no K_4 homeomorph. (Note that this class of graphs contains all trees.) Our main goal is to extend some of the results mentined above with respect to trees. We will use the

* Received by the editors May 3, 1984, and in revised form July 17, 1985.

† Department of Statistics, Tel Aviv University, Tel Aviv 69978, Israel.

above binary decomposition tree to design efficient recursive algorithms for several optimization problems. We will deal mainly with problems which are not affected by adding to a graph edges with infinite length, (e.g. shortest paths problems). (This will enable us to convert a graph with no K_4 homeomorph into an equivalent series parallel graph.) In particular, we produce a linear time algorithm to find the shortest simple paths from a given vertex to all other vertices. We also construct an $O(n^4)$ algorithm to solve the simple plant location problem on an n-vertex graph with no K_4 homeomorph. (The latter is NP-hard for general graphs.)

We also use the binary decomposition tree to prove the existence and find, in linear time, a 2-separator of a graph with no K_4 homeomorph. (A 2-separator of an n-vertex graph $G=(V, E)$ is a pair of vertices $\{u, v\}$ associated with a partition of $V-\{u, v\}$ to sets V_1 and V_2 satisfying (1)-(3) above.)

We mentioned above that the 1-separator of a tree has been used in designing efficient divide and conquer algorithms on trees. We claim that the 2-separator is useful for these purposes on graphs with no K_4 homeomorph. This is demonstrated in § 7, where we study the selection of the kth longest simple path in the set of intervertex distances. This set is of $O(n^2)$ cardinality. However, we will show that the kth largest element in this set can be found in $O(n \log^3 n)$ time (i.e., without explicitly generating the entire set). This result generalizes similar results for tree graphs, [FJ2], [MTZC].

At this point it is worth mentioning the seminal works of Robertson and Seymour, [RS1], [RS2] that extend the concept of a binary decomposition tree, described above, for general graphs. Their results are quite general and therefore are not always most efficient when applied to certain classes of graphs. Specifically, when we apply their results to graphs with no K_4 homeomorph, we get weaker results than those reported above. In particular, their work implies the existence of a 3-separator, instead of the 2-separator result reported above. This is rather crucial for our purposes, since our selection algorithm for the kth longest path will fail if 2-separators are replaced by 3-separators. Also, their general work does not yield linear time algorithms for finding the separators and the binary decomposition tree.

2. Graph-theoretic definitions. A *multigraph* $G=(V, E)$ consists of a finite set of *vertices* V and a finite multiset of *edges* E. Each edge is a pair (u, v) of distinct vertices. If all the edges of G are ordered pairs G is called a *directed* multigraph, and if all its edges are unordered pairs G is an *undirected* multigraph. If E is a set, G is a *graph*, i.e. G contains no parallel edges.

A vertex v is a *cut vertex* of G if the removal of v, and all the edges incident to v, from G results in a disconnected graph. G is *biconnected* if it has no cut vertex. A biconnected *component* of G is a maximal submultigraph of G which is biconnected. A multigraph G contains a subgraph *homeomorphic* to a graph H, if H can be obtained from G by a sequence of the following operations:

(a) remove an edge,

(b) remove an isolated vertex

(c) if a vertex v has degree two, remove v and replace the two distinct edges (u, v) and (v, w) incident to v by an edge (u, w).

This paper focuses on a class of multigraphs, called series-parallel, which are defined as follows:

We say that edges e_1 and e_2 are *in series* if they have a single common vertex which is of degree two. The edges e_1 and e_2 are *parallel* if they have the same set of end vertices. By a *series construction* of an edge $e=(u, v)$ we mean the replacement of e by two edges in series, i.e., by (u, w) and (w, v), such that w is a new vertex. By a *parallel construction* of an edge $e=(u, v)$ we mean the replacement of e by two parallel edges e_1 and e_2, having u and v as their end vertices. By a *series reduction* of two edges in series $e_1=(u, w)$ and $e_2=(w, v)$ we mean the replacement of e_1 and e_2 by a new edge $e=(u, v)$. By a *parallel reduction* of two parallel edges $e_1=(u, v)$ and e_2 we mean the replacement of e_1 and e_2 by a new edge $e=(u, v)$.

A *series-parallel multigraph* (SPM) is a multigraph that can be obtained by a sequence of series and parallel constructions of edges starting from a single edge. Alternatively, G is SPM if it can be reduced to a single edge by a sequence of series and parallel reductions.

3. Basic results on series-parallel multigraphs. The basic theorem on a SPM is due to Dirac [DIR] and Duffin [DUF].

THEOREM 3.1. *Let G be a multigraph. Then each biconnected component of G is series-parallel if and only if G does not contain a subgraph homeomorph to K_4, the complete graph on four vertices.*

The next result shows that a multigraph that does not contain K_4 as a homeomorph, can be augmented by edges to obtain a biconnected SPM.

THEOREM 3.2. *Let G be a connected multigraph that does not contain K_4 as a homeomorph. If G has k biconnected components, then G can be augmented by at most $k-1$ edges such that the resulting multigraph is a biconnected SPM.*

Proof. The proof is by induction on k, the number of biconnected components. Let $G_1, \cdots, G_k$, $k \geqq 2$, be the biconnected components of G. Each pair of biconnected components has at most one vertex in common. Suppose without loss of generality that G_1 and G_2 have a vertex in common, say u. Let u_i be a vertex in G_i, $i=1, 2$, which is adjacent to u (i.e., (u, u_i) is an edge of G_i). Connect u_1 and u_2 by an edge, and call the new multigraph G'. Clearly G_1 and G_2, with the augmented edge, form a biconnected submultigraph which we denote by $G_{1,2}$. The biconnected components of G' are $G_{1,2}$, $G_3, \cdots, G_k$. A multigraph does not contain a K_4 homeomorph if and only if none of its biconnected components does. Thus, it suffices to show that $G_{1,2}$ does not contain a K_4 homeomorph.

Suppose that $G_{1,2}$ contains a K_4 homeomorph, H, and let x, y, z and v be the four vertices of H of degree 3. Assume first that none of G_1 and G_2 contains all of these four vertices. Thus, without loss of generality, let x be in G_1, y be in G_2, and $x \neq u$, $y \neq u$. Since x and y are of degree 3, there exist in H three (intermediate) vertex disjoint paths connecting x and y. But, this is clearly not possible since the connection in G' between G_1 and G_2 is only via u and or u_1. Next assume without loss of generality that x, y, z and v are all vertices of G_1. Since G_1 has no K_4 homeomorph, it follows (without loss of generality) that there is a path, $P(x, y)$, in H connecting x and y, which contains a subpath $P(u, u_1)$, consisting of no edges of G_1. Replacing $P(u, u_1)$ by the edge (u, u_1) of G_1 we obtain from H a K_4 homeomorph H', which is contained in G_1. This is impossible since G_1 has no K_4 homeomorph.

Remark 3.3. Since the biconnected components of a graph can be obtained in linear time, [TAR], the augmentation process described in the proof of Theorem 3.2 can also be performed in linear time.

SPM's can be constructed recursively as follows. Each is associated with two vertices called *terminals.*

(1) A single edge is SPM. Its end vertices are the two terminals.

(2) If G_1 and G_2 are SPM's with terminals x, y and z, w respectively, so are the multigraphs obtained by each of the following compositions.

(2.1.) *Series composition.* Identify a terminal of G_1, say x, with a terminal of G_2, say z. The terminals of the new multigraph are y and w.

(2.2.) *Parallel composition.* Identify one terminal of G_1, say x, with one terminal of G_2, say z, and the second terminal of G_1, y, with w. The terminals of the new multigraph are the two vertices where the identifications occur.

The above recursive construction of an SPM G is represented by a *binary decomposition tree* [VTL]. This is a rooted binary tree T in which each vertex q corresponds to some series-parallel submultigraph of G, denoted by $G(q)$, obtained as follows. Each leaf (tip) of T represents a distinct edge of G. If q is not a leaf then it is either of a series (S) or parallel (P) type. Let q_1 and q_2 be its two sons. $G(q)$ is the submultigraph obtained from $G(q_1)$ and $G(q_2)$ by the respective series or parallel composition. In particular, if r is the root of T $G(r)$ is the multigraph G. Figure 5 of [VTL] illustrates a binary decomposition tree of an SPM. We note that a linear time algorithm that finds the binary decmposition tree of a given SPM is presented in [VTL].

Remark 3.4. Given a SPM G, a binary decomposition tree associates two terminals with G. Suppose that G is also biconnected. If (u, v) is an edge of G it follows [DUF, Corollary 4] that there exists a binary decomposition tree of G that induces the vertices u and v as the two terminals of G. (This tree can be found in linear time by the algorithm of [VTL] that constructs a binary decomposition tree for a directed series-parallel multigraph.)

In the following sections we will utilize the binary decomposition tree representation of a SPM to produce efficient algorithms for several combinatorial problems defined on series-parallel graphs.

4. Finding a 2-separator of a graph with no K_4 homeomorph. In this section we assume that $G=(V, E)$ is a graph, i.e., it has no parallel edges. Suppose that G has n vertices and $n>2$. A set of vertices $X \subseteq V$ is a *k-separator* of G if the following conditions are met: There exist subsets of vertices V_1 and V_2 such that V_1, V_2, X are pairwise disjoint, $V_1 \cup V_2 \cup X = V$, there is no edge $(u, v) \in E$ such that $u \in V_1$ and $v \in V_2$, and $|X| = k$, $|V_i| \leqq \frac{2}{3}n$, $i=1, 2$.

It is well known [KH] that if G is a tree it has a 1-separator. (For example, the centroid of a tree, [GOL], [KH], is a 1-separator.) We will extend this result and prove that a graph that does not contain a K_4 homeomorph has a 2-separator. We will also exhibit a linear time algorithm for finding a 2-separator.

Using Theorem 3.2 we restrict our analysis to biconnected series-parallel graphs. These graphs are planar and have at most $3n-6$ edges. Given such a graph let T be its binary decomposition tree. As noted above each leaf of T corresponds to a distinct edge of T. Thus, T has $m \leqq 3n-6$ leaves and $m-1$ vertices which are of degree greater than one.

Since G is biconnected and $n>2$, there exists a one-to-one mapping of the vertices of G to the set of edges. Such a mapping can be constructed, for example, by considering some spanning tree of G, say $T(G)$. Then, root $T(G)$ at a leaf of $T(G)$, make it an out-tree, and map each vertex to the edge directed into it and the root into an incident edge not included in $T(G)$. Therefore, we now suppose that each vertex of G corresponds to a distinct leaf of T. (Note that when $m>n$ certain leaves of T are not assigned a vertex of G). Each leaf of T which is assigned a vertex of G is given a unit weight while all other vertices of T have zero weight. Call the leaves with unit weight *weighted leaves.* In linear time, [GOL], [KH], we can find a vertex of T (called a *weighted centroid*), say q, such that each of the (at most 3) connnected components of T, obtained by removing q from T, contains at most $n/2$ weighted leaves. Since $n>2$ q itself is not a leaf and it corresponds to a series-parallel subgraph of G, say $G(q)$. See Fig. 1.

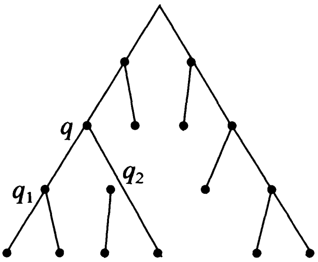


FIG. 1

Let q_1 and q_2 be the two sons of q on T corresponding to the SP subgraphs $G(q_1)$ and $G(q_2)$ respectively. Let u_i, v_i denote the two terminals of $G(q_i)$, $i=1,2$. Let V_i denote the vertex set of $G(q_i)$, $i=1,2$.

Denote by T_i, $i=1,2$, the connected component of T which is obtained by removing q, and contains q_i. T_i contains at most $n/2$ weighted leaves, and $T_1 \cup T_2$ contains at least $n/2$ weighted leaves. Since each weighted leaf in $T_1 \cup T_2$ corresponds to a distinct vertex in $V_1 \cup V_2$, it follows that $|V_1 \cup V_2| \geqq n/2$. Also, each vertex in V_i, $i=1,2$, which is not a terminal corresponds to a distinct weighted leaf in T_i. Thus $|V_i| \leqq n/2+2$, $i=1,2$.

Define $V_3 = V-(V_1 \cup V_2)$. Then $|V_3| \leqq n/2$.

Suppose first that $|V_3| = \max\{|V_1|, |V_2|, |V_3|\}$. Then from $|V_1|+|V_2|+|V_3| \geqq n$ we obtain $n/3 \leqq |V_3| \leqq n/2$. There is no edge of G connecting a vertex in V_3 with a vertex of $G(q)$ which is not one of the two terminals of $G(q)$. (Recall that $G(q)$ is the composition of $G(q_1)$ and $G(q_2)$.) Thus, the two terminals of $G(q)$ constitute a 2-separator of G. It separates V_3 from the remaining vertices of V, i.e., $V-V_3$. Clearly $|V-V_3| \leqq \frac{2}{3}n$ since $|V_3| \geqq n/3$.

Next assume, without loss of generality, that $|V_1| = \max\{|V_1|, |V_2|, |V_3|\}$. Then $n/3 \leqq |V_1| \leqq n/2+2$. There is no edge of G that connects a vertex which is not a terminal of $G(q_1)$ with a vertex in $V-V_1$. Therefore the two terminals of $G(q_1)$ separate the nonterminal vertices of G_1 from $V-V_1$. We have $|V-V_1| \leqq \frac{2}{3}n$ since $|V_1| \geqq n/3$. Also the number of nonterminal vertices of G_1 is $|V_1|-2 \leqq n/2$.

The complexity of the above scheme for locating the 2-separator is governed by the complexity of constructing T, the binary decomposition tree of G, and finding a weighted centroid on it. As mentioned above both take linear time.

We have proved the following.

THEOREM 4.1. *Let $G=(V, E)$ be an n-vertex graph that does not contain K_4 as a homeomorph. There exist two vertices u, v in V and two subsets V_1, $V_2 \subseteq V$, $|V_i| \leqq \frac{2}{3}n$, $i=1,2$, such that $V_1 \cup V_2 \cup \{u, v\} = V$, and there is no edge of G connecting a vertex in V_1 with a vertex in V_2. Furthermore, the vertices, u, v and the sets V_1 and V_2 can be found in linear time.*

Remark 4.2. The above theorem can easily be extended to the weighted version where each vertex v in V is associated with a positive weight w_v. In this case the condition $|V_i| \leqq \frac{2}{3}n$ is replaced by $\sum_{v \in V_i} w_v \leqq \frac{2}{3} \sum_{v \in V} w_v$.

Remark 4.3. From the above discussion it follows that the 2-separators which we find consist of a pair of vertices $\{u, v\}$ that are terminals of some series-parallel subgraph. Therefore if we add the edge (u, v) to the underlying graph $G=(V, E)$ the resulting graph will also not contain K_4 as a homeomorph.

5. Linear time algorithms for finding shortest paths on series-parallel graphs. We have already noted above that a binary decomposition tree of a series-parallel graph can be generated in linear time [VTL]. With the aid of this tree one can design efficient algorithms for solving a variety of combinatorial problems defined on the underlying series-parallel graph. For example, linear time algorithms for finding a minimal vertex cover, a minimal dominating set or a Hamiltonian cycle can easily be constructed, utilizing the decomposition tree.

In this section we demonstrate the use of the decomposition tree to solve edge-weighted problems on graphs that do not contain K_4 as a homeomorph. Although we will focus on the problem of finding shortest (simple) paths, the approach can easily be modified to solve other problems like the matching problem and the travelling salesman problem. In particular, since the dual of a series-parallel graph is also series-parallel, the method described in [HAS] combined with the algorithm of this section yields a linear time algorithm for the maximum flow problem on series-parallel graphs.

To this end let $G=(V, E)$ be a connected directed graph with $|V|=n$, that does not contain K_4 as a homeomorph. Each (directed) edge e in E has a "length" (weight) w_e (which is not necessarily nonnegative). For each pair of vertices u and v define $d(u, v)$ to be length of a shortest simple directed path from u to v. Given a vertex s in V, we wish to find the distances from s to all vertices v in V. This problem is recognized as the *single source shortest paths problem.*

If the graph may contain negative cycles (we do not exclude this possibility), the problem is NP-hard for general planar graphs. If $w_e \geqq 0$ for each edge, then this problem is solvable for any n-vertex planar graph in $O(n\sqrt{\log n})$ time, [FRED]. (In particular, this bound is applicable to our class of graphs, since these graphs are planar.) However, we will show that the single source shortest paths problem is solvable in linear time for graphs with no K_4 homeomorph.

Without loss of generality we assume that G is biconnected and series-parallel, since otherwise we may apply the procedure described in the proof of Theorem 3.2, and add to G edges with infinite length. We consider (as noted in Remark 3.4), a decomposition tree T, that induces the vertex s (from which the distances $d(s, v)$, $v \in V$, are sought for) as one of the terminals of G.

We now give a brief overview of the algorithm, which consists of two phases. In the first phase we start with the leaves of T and recursively compute for each subgraph of G, $G(q)$, (that corresponds to a vertex q of T), the distances, restricted to $G(q)$, $d(q, i(q), j(q))$ and $d(q, j(q), i(q))$ between its two terminals $i(q)$ and $j(q)$. The recursive equations are rather straightforward and therefore they are omitted.

In the second phase we start with the root of the tree (corresponding to the graph G), and recursively compute the distances in G from s to the terminals of each subgraph $G(q)$. Note that each vertex of G is a terminal of at least some subgraph $G(q)$. Therefore, at the end of the second phase we will have the distances $d(s, v)$ for all vertices in V. The second phase uses the following auxiliary distance function. If x is a terminal of $G(q)$, let $d'(q, s, x)$ denote the length of a directed shortest simple path from s to x that does not contain any edge of $G(q)$. Starting with the root of T, we recursively compute $d'(q, s, x)$ and $d(s, x)$. The recursive formulae are again quite clear. For example, suppose that $G(q)$ is generated from $G(q_1)$ and $G(q_2)$ by a series composition with, say, $i(q)=i(q_1)$ and $j(q)=j(q_2)$, (i.e., $j(q_1)=i(q_2)$). Then,

$$d'(q_1, s, i(q_1)) = d'(q, s, i(q)),$$

$$d'(q_1, s, j(q_1)) = d'(q, s, j(q)) + d(q_2, j(q_2), i(q_2)),$$

$$d(s, i(q_1)) = d(s, i(q)),$$

and

$$d(s, j(q_1)) = \text{Min}\,[(d'(q, s, i(q)) + d(q_1, i(q_1), j(q_1))), \qquad (d'(q, s, j(q)) + d(q_2, j(q_2), i(q_2)))].$$

For the benefit of brevity we omit the recursive equations for the other cases.

We claim that the above algorithm solves the single source shortest paths problem in linear time. The linear complexity follows from the facts that the decomposition tree has $O(n)$ vertices, and that each recursive equation used in either one of the two phases can be computed in constant time.

6. The uncapacitated plant location problem. The uncapacitated (or simple) plant location problem is defined on a n-vertex undirected graph $G = (V, E)$. Each edge e in E is assumed to have a positive length w_e. Each vertex v in V is associated with a fixed cost c_v (the set up cost of a facility at v), and a transportation cost function $f_v(\cdot)$, which is a nondecreasing function of its argument. Each vertex is viewed as both a potential site for a service center (facility) as well as a customer. The uncapacitated plant location problem is to locate facilities at vertices so as to minimize the sum of the setup costs and the transportation costs. Since the transportation cost functions are nondecreasing each customer will use the closest facility to him. Formally the problem is to find a subset of vertices $S \subseteq V$ that minimizes

$$\sum_{v \in S} c_v + \sum_{v \in V} f_v(d(v, S)),$$

where

$$d(v, S) = \underset{u \in S}{\text{Min}}\; d(v, u).$$

This problem is NP-hard on general planar graphs. However, $O(n^2)$ algorithms are given in [KT], for the case when G is a tree. In this section we will present an $O(n^4)$ algorithm solving the problem for graphs containing no K_4 as a homeomorph.

As in the previous section, the procedure will be based on a recursive approach on the corresponding binary decomposition tree.

Since we can add to G edges with infinite length, we use Theorem 3.2 and assume without loss of generality that G is a biconnected series-parallel graph and T is its binary decomposition tree. We will use the terminology that customer v is served by facility u if u is the closest established facility to v.

Let q be a vertex of T, $G(q)$ its respective series-parallel subgraph of G and $i(q)$ and $j(q)$ the two terminals of $G(q)$. For each pair of vertices u, v in V we define $h(q, u, v)$ to be the solution value of the uncapacitated plant location problem restricted to the subgraph $G(q)$, with the constraint that $i(q)$ is served by u and $j(q)$ is served by v. u and v are not restricted to be in $G(q)$. Formally, if we let $V(q)$ denote the vertex set of $G(q)$, $h(q, u, v)$ is defined by

$$h(q, u, v) = \underset{S \subseteq V}{\text{Min}} \left\{ \sum_{w \in S} c_w + \sum_{w \in V(q)} f_w(d(w, S)) \right\}$$

$$\text{s.t. } i(q) \text{ is served by } u \text{ and } j(q) \text{ is served by } v.$$

The solution to the uncapacitated plant location problem on G is, therefore, given by $\text{Min}_{u,v \in V}\, h(r, u, v)$, where r is the root of the decomposition tree. The functions $h(q, \cdot, \cdot)$ will be evaluated, recursively, starting with the leaves of T. Suppose that $G(q)$ is a composition of $G(q_1)$ and $G(q_2)$. For a series composition with, say, $i(q) = i(q_1)$,

$j(q)=j(q_2)$, (i.e., $j(q_1)=i(q_1)$) we obtain

$$h(q, u, v)=\operatorname*{Min}_{w\in V(q)\cup\{u,v\}}\{h(q_1, u, w)+h(q_2, w, v)-f_{j(q_1)}(d(j(q_1), w))-c_w-c_u\delta(u, v, w)\},$$

where $\delta(u, v, w)$ is equal to 1 if $u=v=w$ and 0 otherwise. For a parallel composition with $i(q)=i(q_1)=i(q_2)$ and $j(q)=j(q_1)=j(q_2)$ we have $h(q, u, v)=h(q_1, u, v)+h(q_2, u, v)-f_{i(q)}(d(i(q), u))-f_{j(q)}(d(j(q), v))-c_u-c_v\varepsilon(u, v)$, where $\varepsilon(u, v)$ is equal to 1 if $u\neq v$ and 0 otherwise.

To compute the complexity of this recursive procedure, we assume that it takes a constant time to compute any cost function $f_v(x)$ at a given value of x. To compute these functions we need the distances between all pairs of vertices in G. If we use the algorithm given in § 5, this task can be performed in $O(n^2)$ time. It follows from the above recursive formulae that the complexity bound of the proecure is $O(kn^3)$, where $k\geqq 1$ is the number of vertices of the decomposition tree of the series type.

7. Finding the kth longest distance on series parallel graphs. Let $G=(V, E)$ be an n-vertex directed graph with nonnegative edge-lengths $w_e\geqq 0$, $e\in E$, and consider the (multi) set of intervertex distances

$$S=\{d(u, v)\,|\,u, v\in V, u\neq v\}.$$

($d(u, v)$ is the length of a shortest simple directed path from u to v.)

The cardinality of S is $O(n^2)$, and therefore if S is given explicitly, the kth largest element in S can be found in $O(n^2)$ time by a standard algorithm [AHU]. For general graphs no algorithm is known for finding even the largest element in S in time which is of a lower order than the time required to compute S explicitly. (The best time bound known is $O(n^3(\log\log n/\log n)^{1/3})$ [FRE].) When G is a tree, S can be given a succinct representation that requires only $O(n \log n)$ space. The construction of this representation takes $O(n \log n)$ time, and it enables the selection of the kth largest element in S in $O(n \log n)$ time, [FJ2], [MTZC], i.e., the time is sublinear in the size of S. (We note that for trees the largest element in S can be found in linear time, [HAN].)

As discussed in [FJ2], [MTZC], the ability to select efficiently in the set S gives rise to efficient algorithms for a variety of center location problems.

In this section we extend the results of [FJ2], [MTZC] and show that if G does not contain K_4 as a homeomorph, the kth largest element in S can be selected in $O(n \log^3 n)$ time, using $O(n \log n)$ space.

The algorithm is a divide-and-conquer scheme, which is based extensively on the linear time algorithms of §§ 4 and 5 to compute 2-separators and distances.

The first phase of the algorithm deals with the succinct representation of S. We partition S into $\bar{m}=O(n \log n)$ subsets such that the k_jth largest element in the jth subset, for all $j=1, \cdots, \bar{m}$, can be found in a total time of $O(\bar{m} \log n)=O(n \log^2 n)$.

The partitioning phase starts by finding a 2-separator of G consisting of a pair of vertices u, v. Let V_1, V_2 be the two subsets of V separated by u, v (Theorem 4.1). Without loss of generality we assume that the three sets V_1 and V_2 and $\{u, v\}$ are mutually disjoint.

At this stage we partition only the subsets of S corresponding to distances from vertices in V_1 to vertices in V_2 and from vertices in V_2 to vertices in V_1. (We then proceed recursively by decomposing $V_1\cup\{u, v\}$ and $V_2\cup\{u, v\}$.)

For any pair of vertices $x\in V_1$ and $y\in V_2$, any simple path connecting the two must contain at least one of the separators u and v. Moreover, since all edge lengths are nonnegative, we have

$$d(x, y) = \text{Min}\,\{d(x, u) + d(u, y), d(x, v) + d(v, y)\},$$
$$d(y, x) = \text{Min}\,\{d(y, u) + d(u, x), d(y, v) + d(v, x)\}.$$

For any disjoint subsets of vertices $X, Y \subseteq V$ we will define

$$S(X, Y) = \{d(x, y) : x \in X, y \in Y\}.$$

Focusing first on $S(V_1, V_2)$ we partition it to $S^-(V_1, V_2)$ and $S^+(V_1, V_2)$, where

$$S^-(V_1, V_2) = \{d(x, u) + d(u, y) | x \in V_1, y \in V_2, d(x, u) - d(x, v) + d(u, y) - d(v, y) \leqq 0\},$$
$$S^+(V_1, V_2) = \{d(x, v) + d(v, y) | x \in V_1, y \in V_2, d(x, u) - d(x, v) + d(u, y) - d(v, y) > 0\}.$$

Defining $a_x = d(x, u)$, $c_x = d(x, u) - d(x, v)$, $x \in V_1$ and $b_y = d(u, y)$, $d_y = d(u, y) - d(v, y)$, $y \in V_2$, we note that $S^-(V_1, V_2)$ (and similarly $S^+(V_1, V_2)$) are both of the nature of the sets described in Appendix 1. The sequences $\{a_x\}$, $\{c_x\}$, $x \in V_1$, and $\{b_y\}$, $\{d_y\}$, $y \in V_2$ can be constructed in linear time if we find the 2-separator $\{u, v\}$ by the algorithm in § 4 and the distances from (and to) u and v by the algorithm in § 5.

Using Appendix 1 and noting that $|V_i| = O(n)$, $i = 1, 2$, we conclude that $S(V_1, V_2)$ can be partitioned into $m' = O(n)$ lists and stored in a data structure (of $O(n)$ space), on which we can select the k_jth largest element in the jth list for all $j = 1, \cdots, m'$ in a total time of $O(n \log n)$.

Clearly the same partitioning will be applied to the set $S(V_2, V_1)$. From Appendix 1 we see that the complexity bound of this stage of the recursion is $O(n \log n)$.

We now have to proceed with the distances within the set of vertices $V_1 \cup \{u, v\}$ (and similarly the distances within $V_2 \cup \{u, v\}$).

Let G_i, $i = 1, 2$ denote the subgraph of G induced by the vertices $V_i \cup \{u, v\}$. To compute the distances within, say, $V_1 \cup \{u, v\}$, we may not restrict our discusion to G_1, since the shortest path connecting u to v (or v to u) may not be included in G_1. However, this difficulty can easily be resolved by adding to G_i, $i = 1, 2$, the (directed) edges (u, v) and (v, u) with edge-lengths $d(u, v)$ and $d(v, u)$ respectively. (Note that $d(u, v)$ and $d(v, u)$ have already been computed earlier). To proceed recursively with G_i, $i = 1, 2$, we need the property that G_i, $i = 1, 2$, with the above augmented edges does not contain K_4 as a homeomorph. Indeed, this is guaranteed by Remark 4.3. To make sure that each pair of vertices is taken into account precisely once, when we deal with the set of distances within $V_2 \cup \{u, v\}$ we will not record the distances between u and v.

We now estimate the total number of lists created during the entire partitioning process. Let $f(n)$, denote the maximum number of such lists created in such a partition of a graph with n vertices. Then

$$f(n) \leqq c_1 n + f(n_1) + f(n_2),$$

where $n_i = |V_i \cup \{u, v\}| \leqq \frac{2}{3}n + 2$, $i = 1, 2$, and $n_1 + n_2 = n + 2$. Therefore, $f(n) = O(n \log n)$.

To estimate the total space needed for the process, $g(n)$, we note that $g(n)$ satisfies

$$g(n) \leqq c_2 n + g(n_1) + g(n_2)$$

with n_i as above, and thus $g(n) = O(n \log n)$. Finally, the running time, $T(n)$, of the partitioning phase satisfies

$$T(n) \leqq c_3 n \log n + T(n_1) + T(n_2)$$

where $n_i \leqq \frac{2}{3}n + 2$ $i = 1, 2$, and $n_1 + n_2 = n + 2$. It thus follows that $T(n) = O(n \log^2 n)$.

Next we turn to the selection phase. At this point S is represented by $\bar{m} = O(n \log n)$ lists (subsets), say, $S_1, \cdots, S_{\bar{m}}$. These lists are maintained in a data structure on which the total time to select the k_jth largest element from S_j for all $j = 1, \cdots, \bar{m}$ is $O(\bar{m} \log n)$, (Appendix 1). Since each list is of cardinality $O(n)$, we can use the algorithm of [FJ1] to select the k-th largest element (i.e., kth longest distance) in S in $O(n \log^3 n)$ time. (Note that the time bound $O(\bar{m} \log n)$ mentioned in [FJ1] is based on the assumption that the total time to select the k_jth largest element from S_j for all $j = 1, \cdots, \bar{m}$ is $O(\bar{m})$. Since in our case the latter is replaced by $O(\bar{m} \log n)$, our total time bound is $O(\bar{m} \log^2 n) = O(n \log^3 n)$.)

Finally we note that if we only wish to find the largest element in S, i.e., the *diameter* of the graph, we do not need the first phase. In fact, the diameter can be found recursively by finding at each iteration, the largest element in $S(V_1, V_2) \cup S(V_2, V_1)$. The latter can be found in $O(n)$ time, thus yielding an $O(n \log n)$ total time for the scheme that finds the diameter.

Appendix 1. Representation of and selection in the set $R = \{a_i + b_j | c_i + d_j \leqq 0, i = 1, \cdots, n, j = 1, \cdots, m\}$. Suppose that we are given four sequences of real numbers, $\{a_i\}$, $i = 1, \cdots, n$, $\{c_i\}$, $i = 1, \cdots, n$, $\{b_j\}$, $j = 1, \cdots, m$, $\{d_j\}$, $j = 1, \cdots, m$. Define the (multi-) set R by

$$R = \{a_i + b_j | c_i + d_j \leqq 0, i = 1, \cdots, n, j = 1, \cdots, m\}. \tag{A.1}$$

Our goal is to provide a partition of the multiset R, (whose cardinality is $O(mn)$) into m lists $R'_1, \cdots, R'_m$, and a data structure (requiring a total space of $O(m+n)$), for storing these lists, on which the following operations can be implemented: Given integers $k_1, \cdots, k_m$, select the k_jth largest element in R'_j for all $j = 1, \cdots, m$, in $O((m+n) \log n)$ total time. (The total time consumed in constructing this data structure will be shown to be $O(m \log m + n \log n)$.)

For each j, $j = 1, \cdots, m$, define the set $R_j = \{a_i | c_i + d_j \leqq 0, i = 1, \cdots, n\}$. Sort the sequences $\{c_i\}$, $i = 1, \cdots, n$, and $\{d_j\}$, $j = 1, \cdots, m$, and assume without loss of generality that $c_1 \leqq c_2 \cdots \leqq c_n$, and $d_1 \leqq d_2 \leqq \cdots \leqq d_m$. Next, define $I(j) = \{i | c_i + d_j \leqq 0\}$, $j = 1, \cdots, m$, to note that

$$I(1) \supseteq I(2) \supseteq \cdots \supseteq I(m). \tag{A.2}$$

Rewriting R_j, $j = 1, \cdots, m$, as $R_j = \{a_i | i \in I(j)\}$, define h_j, $j = 1, \cdots, m$, to be the difference between adjacent sets, i.e.,

$$h_j = |I(j)| - |I(j+1)|.$$

The subset R_{j+1} is obtained from R_j by deletion of h_j elements. Thus using a 2-3 tree data structure, [AHU], we can delete the respective elements, and find the k_jth largest element in R_j for all $j = 1, \cdots, m$, in a total time of $O((\sum_{j=1}^{m-1} h_j) \log n + m \log n)$. By (A.2) this bound is $O((m+n) \log n)$.

The total effort to construct this data structure is dominated by the sorting of the sequences, $\{c_i\}$, $i = 1, \cdots, n$ and $\{d_j\}$, $j = 1, \cdots, m$, i.e., $O(n \log n + m \log m)$.

The multiset R, defined by (A.1), can now be represented as the union of the following m subsets, $R'_1, \cdots, R'_m$, where $R'_j = \{b_j\} + R_j$, $j = 1, \cdots, m$. Using the above we conclude that the k_jth largest element in R'_j for all $j = 1, \cdots, m$, can be found in $O((m+n) \log n)$ total time by selecting the elements in R_j, $j = 1, \cdots, m$ and adding the appropriate b_j value.

REFERENCES

[AHU] A. V. AHO, J. E. HOPCROFT AND J. D. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[BBT] W. W. BEIN, P. BRUCKER AND A. TAMIR, *Minimum cost flow algorithms for series parallel networks*, Discrete Appl. Math., 10 (1985), pp. 117-124.

[COL] R. COLE, *Searching and storing similar lists*, Technical Report 88, Dept. Computer Science, Courant Institute of Mathematical Sciences, New York Univ. New York, Oct., 1983.

[DIR] G. A. DIRAC, *A property of 4-chromatic graphs and some remarks on critical graphs*, J. London Math. Soc., 17 (1952), pp. 85-92.

[DUF] R. J. DUFFIN, *Topology of series parallel graphs*, J. Math. Anal. Appl., 10 (1965), pp. 303-318.

[FRE] M. FREDMAN, *New bounds on the complexity of the shortest path problem*, SIAM J. Comput., 5 (1976), pp. 83-89.

[FRED] G. N. FREDERICKSON, *Shortest path problems in planar graphs*, in Proc. 24th IEEE Symposium on Foundations of Computer Science, Tucson, AZ, Nov. 1983, pp. 242-247.

[FJ1] G. N. FREDERICKSON AND D. B. JOHNSON, *The complexity of selection and ranking in $X+Y$ anc! matrices with sorted columns*, J. Comput. System Sci., 24 (1982), pp. 197-208.

[FJ2] ———, *Finding kth paths and p-centers by generating and searching good data structures*, J. Algorithms, 4 (1983), pp. 61-80.

[GOL] A. J. GOLDMAN, *Optimal center location in simple networks*, Transport. Sci., 5 (1971), pp. 212-221.

[HAN] G. Y. HANDLER, *Minimax location of a facility in an undirected tree graph*, Transport. Sci., 7 (1973), pp. 287-293.

[HAS] R. HASSIN, *Maximum flow in (s, t) planar networks*, Inform. Proc. Letters, 13 (1981), p. 107.

[KH] O. KARIV AND S. L. HAKIMI, *An algorithmic approach to network location problems, Part* I. *The p-centers*, SIAM J. Appl. Math., 37 (1979), pp. 513-538.

[KT] A. KOLEN AND A. TAMIR, *Covering problems*, to appear in Discrete Location Theory, L. R. Francis and P. B. Mirchandani, eds., John Wiley, New York, 1986.

[MTZC] N. MEGIDDO, A. TAMIR, E. ZEMEL AND R. CHANDRASEKARAN, *An $O(n \log^2 n)$ algorithm for the kth longest path in a tree with applications to location problems*, SIAM J. Comput., 10 (1981), pp. 328-337.

[RS1] N. ROBERTSON AND P. D. SEYMOUR, *Graph minors* I: *Excluding a forest*, J. Comb. Theory B, 35 (1983), pp. 39-61.

[RS2] ———, *Graph minors* II: *Algorithmic aspects of tree width*, submitted for publication.

[ST] G. STEINER, *A compact labeling scheme for series parallel graphs*, Discrete Appl. Math., 11 (1985), pp. 281-297.

[TAR] R. E. TARJAN, *Depth-first search and linear graph algorithms*, SIAM J. Comput., 1 (1972), pp. 146-160.

[TNN] N. TAKAMIZAWA, T. NISHIZEKI AND N. SAITO, *Linear-time computability of combinatorial problems on series parallel graphs*, J. ACM, 29 (1982), pp. 623-641.

[VTL] J. VALDES, R. E. TARJAN AND E. L. LAWLER, *The recognition of series parallel diagraphs*, SIAM J. Comput., 11 (1982), pp. 298-313.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

007

CONVERGENT ITERATIONS FOR COMPUTING STATIONARY DISTRIBUTIONS OF MARKOV CHAINS*

G. P. BARKER† AND R. J. PLEMMONS‡

Abstract. Classical iterative schemes such as the Gauss-Seidel method and its variations constitute powerful tools for computing stationary distribution vectors for large-scale Markov process, such as those arising in queueing network analysis. The coefficient matrix A in these processes in a Q-matrix, i.e., a singular irreducible M-matrix with zero column sums and, unlike the nonsingular case, the classical iterations for A do not always converge. The purpose of this paper is to survey the recent literature and to analyze the behavior of these methods completely in terms of the graph structure of A. The results given here hold under somewhat weaker assumptions on A.

Key words. iterative methods, Markov chains, singular M-matrices, queueing networks, sparse matrices

AMS(MOS) subject classifications. 65F10, 15A06, 15A48, 60K20

1. Introduction.

1.1. Background. In this paper we primarily consider $n \times n$ irreducible matrices

$$A = (a_{ij}) \quad \text{with } a_{ij} \leqq 0 \text{ for all } i \neq j \text{ and with } \sum_{i=1}^{n} a_{ij} = 0,\ 1 \leqq j \leqq n.$$

Adopting the terminology used in Rose [1984], and elsewhere, we call such matrices *Q-matrices.* These matrices arise in several areas, including the analysis of queueing networks, from whence the term Q-matrix orginated, (see, e.g., Kaufman [1983]), in the analysis of compartmental models in the biological sciences (see, e.g., Funderlic and Mankin [1981]), and in the input-output analysis of economic models (see, e.g., Berman and Plemmons [1979, Chap. 9]). Q-matrices form an important subclass of the widely studied class of M-matrices and thus they possess several useful properties described, e.g., in Berman and Plemmons [1979, Chap. 6].

Let A be a Q-matrix. In many of the applications listed above the stationary distribution vector p for an underlying Markov process associated with A is of primary importance. In particular the solution of the homogeneous system of linear equations

$$Ax = 0 \tag{1.1}$$

where A is a Q-matrix is of interest. Our purpose is to compute the unique stationary probability distribution vector $p = (p_i)$, $p_i > 0$, $\sum_{i=1}^{n} p_i = 1$, which solves (1.1). Here A is considered to be the transfer rate matrix for a finite ergodic process. The evaluation of the stationary probability vector p of such a process, defined by its transition probability matrix Q, is a classical problem in the modeling and the performance analysis of computer systems, of data communication networks, or of telephone exchange systems. Here p is the left positive eigenvector of Q and thus $Ap = 0$ where $A = I - Q^T$.

Various problems related to the computation of p have drawn considerable attention recently. Iterative methods have been studied, e.g., by Buoni [1986], Courtois

* Received by the editors July 28, 1983, and in revised form August 6, 1985.

† Department of Mathematics, University of Missouri at Kansas City, Kansas City, Missouri 64110.

‡ Departments of Computer Science and Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205. The research of this author was supported in part by the Air Force Office of Scientific Research under grant AFOSR-83-0255, and by the National Science Foundation under grant DMS-82-19500-A02.

and Semal [1985], Funderlic and Plemmons [1984], Kaufman [1983], Koury, McAllister and Stewart [1984], Lubacheuski and Mitra [1985], Rose [1984], Schneider [1984], and Stewart, Stewart and McAllister [1985]. In addition, the sensitivity analysis of p in terms of perturbations of A has recently been studied, e.g., by Barlow [1985], Funderlic and Meyer [1985] and Golub and Meyer [1985].

In many applications, the Q-matrix A is quite large and sparse. In fact, problems with 100,000 or more equations are not uncommon, especially in queueing network applications, as described by Kaufman [1983]. In this regard, variations of the power method for Q and methods based upon Jacobi and Gauss-Seidel type splittings of $A = I - Q^T$ constitute powerful tools for computing the stationary distributions. We point particularly to the recent work of Kaufman [1983] in which point Gauss-Seidel type methods are investigated, to the work Lubacheuski and Mitra [1985] in which a chaotic asynchronous variation of the power method is investigated for parallel processors and to the work of Stewart, Stewart and McAllister [1985], in which a two-stage iteration based upon the Gauss-Seidel method is considered. It is these papers that motivated our work described herein.

Unlike the nonsingular M-matrix case, the classical iterative methods (the power, Jacobi and Gauss-Seidel methods) applied to the computation of p do not always converge. In this paper we survey and clarify the recent literature on convergence results for these computations and summarize, by applying elementary properties of nonnegative matrices, necessary and sufficient conditions for convergence in terms of the graph structure of Q or A. Our results are given for point iterations. Analogous results can be stated for more general block iterative schemes. Several numerical examples are given in order to illustrate the variety of convergence situations that can occur.

1.2. Notation and conventions. For a real $n \times n$ matrix $B = (b_{ij})$ the *directed graph* $\Gamma(B)$ of B is the graph with vertices $1, 2, \cdots, n$ and edges (i, j) for $b_{ij} \neq 0$. As in Rose [1984], we say that vertex v_1 is *adjacent to* vertex v_2 if (v_1, v_2) is an edge of $\Gamma(B)$. A *path* of length $\lambda - 1$ is an ordered set $p = (v_1, \cdots, v_\lambda)$ such that for each i, v_i is adjacent to v_{i+1}. If $v_\lambda = v_1$, then p is called a *closed path*. A closed path for which $v_1, \cdots, v_{\lambda-1}$ are distinct is a *cycle* (a circuit in Schneider [1984]). A *monotone path* is a path $p = (v_1, \cdots, v_\lambda)$ for which either $v_1 < \cdots < v_\lambda$ (*monotone increasing*) or $v_1 > \cdots > v_\lambda$ (*monotone decreasing*). By a slight abuse of the language, a closed path $p = (v_1, \cdots, v_\lambda, v_1)$ will be called a *monotone cycle* if p is a cycle and $p_1 = (v_1, \cdots, v_\lambda)$ is monotone with $v_\lambda < v_1$ if p_1 is decreasing and $v_\lambda > v_1$ if p_1 is increasing.

For an $n \times n$ matrix B, the graph $\Gamma(B)$ is called *strongly connected* if every pair of vertices is connected by a path in $\Gamma(B)$. In this case B is said to be *irreducible.* Equivalently, B is irreducible if and only if there is no permutation matrix P for which

$$P^TBP = \begin{bmatrix} B_1 & 0 \\ B_2 & B_3 \end{bmatrix}$$

where B_1 and B_3 are square submatrices of B.

The following terminology and elementary facts about nonnegative matrices can be found in, e.g., Berman and Plemmons [1979]. A nonnegative matrix B is called *primitive* if some power of B consists only of positive entries. It follows that B is primitive if and only if B is irreducible and the greatest common divisor of the lengths of all the cycles in $\Gamma(B)$ is 1. In this case $\rho(B)$, the spectral radius of B, is a simple eigenvalue of B. Next, an $n \times n$ matrix A is called a *singular M-matrix* if it can be written as

$$A = \rho(B)I - B$$

where B is nonnegative. In particular, if Q is the transition probability matrix for an ergodic (irreducible) Markov chain, then the Q-matrix $A = I - Q^T$ is an irreducible singular M-matrix. For such matrices we are concerned in this paper with *regular splittings*, i.e., splittings of the form $A = M - N$, M^{-1} and N nonnegative, and the resulting iterative schemes

$$Mx^{(m+1)} = Nx^{(m)}, \qquad m = 0, 1, \cdots. \tag{1.2}$$

It is well known (see, e.g., Neumann and Plemmons [1978]) that for regular splittings of a singular irreducible M-matrix A, the general iterative scheme (1.2) converges if and only if 1 is the only eigenvalue of the iteration matrix $M^{-1}N$ on the unit circle.

In the next section we first review and summarize the necessary and sufficient conditions for convergence of the power method and iterations based upon the Jacobi splitting of A. We will be mainly concerned, however, with convergence criteria for the Gauss-Seidel method. The results here are based upon recent work of Rose [1984] and Schneider [1984] and are given for point Gauss-Seidel. However, similar results can be stated for the more general block Gauss-Seidel splitting, based upon a block partitioned form of A (see Courtois and Semal [1985] and Rose [1984] for discussions of block iterations for Q-matrices).

2. Convergence criteria.

2.1. The power and Jacobi methods. First, suppose that Q is the transition probability matrix for an ergodic Markov chain. Then the (point) *power method*

$$x^{(m+1)} = Q^T x^{(m)}, \qquad m = 0, 1, \cdots \tag{2.1}$$

results from a regular splitting of A into $A = I - Q^T$.

Next, let A denote a singular irreducible M-matrix. By a (point) *splitting* of A we mean a splitting

$$A = D - L - U \tag{2.2}$$

where D, L and U are the diagonal, strictly lower triangular and strictly upper triangular submatrices of A, respectively. Then the *Jacobi method*

$$Dx^{(m+1)} = (L + U)x^{(m)}, \qquad m = 0, 1, \cdots \tag{2.3}$$

results from a regular splitting of A, since D is a positive diagonal matrix. Note that if we replace A by $D^{-1}A$, then (2.2) corresponds to the power method for $L + U$, although $L + U$ need not be stochastic. The following well-known proposition reviews conditions on the graph structure of A alone in order that the Jacobi method converge. By a *proper cycle* of a matrix we mean a cycle of length at least 2.

PROPOSITION 1. *Let A be an irreducible singular M-matrix. Then the following statements are equivalent.*

1. *The Jacobi method* (2.3) *for A converges for each $x^{(0)}$.*
2. *$D - A$ is primitive.*
3. *The greatest common divisor of the lengths of all proper cycles in A is one.*

Proof. From § 1, it follows that (2.3) converges if and only if $D - A = L + U$ is primitive, since $\rho(L + U) = 1$ and $L + U$ is irreducible. An application of Berman and Plemmons [1979, p. 35], or Varga [1962, p. 69], establishes the equivalence of 2 and 3. □

The proof of the following proposition concerning the power method for computing stationary distributions is analogous to the proof of Proposition 1.

PROPOSITION 2. *Let $A = I - Q^T$, where Q is the transition probability matrix for an ergodic Markov chain. Then the following statements are equivalent.*

1. *The power method* (2.1) *for computing the stationary distribution vector converges for each $x^{(0)}$.*
2. *$I - A^T = Q$ is primitive.*
3. *The greatest common divisor of the lengths of all cycles in A is one.*

Observe that from part 3 of Proposition 2, it follows that the power method must converge whenever Q has at least one nonzero diagonal entry, for then $A = I - Q^T$ must have a cycle of length one.

Also note that the power method must converge for Q whenever the Jacobi method converges for the matrix $A = I - Q^T$. However, the converse of this statement is not true, as the following simple example shows.

Let

$$Q = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Then Q is primitive since it is irreducible with a nonzero diagonal entry, and thus the power method converges. However, the Jacobi method for

$$A = I - Q^T = \begin{bmatrix} \frac{1}{2} & -1 & 0 \\ 0 & 1 & -1 \\ -\frac{1}{2} & 0 & 1 \end{bmatrix}$$

does not converge since A has one proper cycle and it has length 3.

2.2. The Gauss–Seidel Method. Let $A = D - L - U$ be a singular irreducible M-matrix. Then there are two possible Gauss-Seidel iterations for solving homogeneous system $Ax = 0$. These are

$$(D - L)x^{(m+1)} = Ux^{(m)}, \qquad m = 0, 1, \cdots, \tag{2.4}$$

with the iteration matrix

$$T_L = (D - L)^{-1}U, \tag{2.5}$$

and

$$(D - U)x^{(m+1)} = Lx^{(m)}, \qquad m = 0, 1, \cdots, \tag{2.6}$$

with the iteration matrix

$$T_U = (D - U)^{-1}L. \tag{2.7}$$

The iterations (2.4) and (2.6) will be called *forward* and *backward Gauss-Seidel*, respectively. It is well known that these iterations may not converge for an arbitrary Q-matrix A (see the examples given later). If $A = I - Q^T$, Q stochastic, and convergence results then the stationary distribution vector p can be obtained as the limit by scaling $x^{(k)}$ so that $\sum_{i=1}^{n} x_i^{(k)} = 1$.

The iterations (2.4) and (2.6) converge for each $x^{(0)}$ if and only if the powers of the iteration matrices (2.5) and (2.7), respectively, converge. It follows then that the nonnegative matrices T_L and T_U converge if and only if 1 is the only eigenvalue of modulus 1. From Rose [1984] or Schneider [1984] we know that there exist permutation matrices P_L and P_U such that

$$P_L^T T_L P_L = \left[\begin{array}{c|c} 0 & * \\ \hline 0 & H_L \end{array}\right] \tag{2.8}$$

and

$$P_U^T T_U P_U = \left[\begin{array}{c|c} H_U & 0 \\ \hline * & 0 \end{array}\right], \tag{2.9}$$

where H_L and H_U are irreducible nonnegative matrices. This establishes the following elementary observation on the convergence of Gauss-Seidel iterations in terms of H_L and H_U.

LEMMA 1. *The forward (backward) Gauss-Seidel iteration scheme converges for each $x^{(0)}$ if and only if the matrix H_L given by* (2.8) (*H_U given by* (2.9)) *is primitive, that is, some power is positive.*

In order to characterize convergence in terms of the graph structure of A alone we need some special notation. Let α be any cycle in the graph $\Gamma(A)$ and let u_α and l_α *denote the number of edges of* α *which lie in the graph* $\Gamma(U)$ *and* $\Gamma(L)$, *respectively.* Our main result follows. Here "*gcd*" is an abbreviation for the "greatest common divisor".

THEOREM 1. *Let A be a singular irreducible M-matrix. Then*:

1. *The forward Gauss-Seidel method* (2.4) *converges for each $x^{(0)}$ if and only if*

$$\gcd\{u_\alpha\colon \alpha \textit{ is a cycle in } \Gamma(A)\} = 1.$$

2. *The backward Gauss-Seidel method* (2.6) *converges for each $x^{(0)}$ if and only if*

$$\gcd\{l_\alpha\colon \alpha \textit{ is a cycle in } \Gamma(A)\} = 1.$$

Proof. We consider T_L given by (2.5) since the argument will apply to T_U given by (2.7) mutatis mutandis. We wish to apply Schneider [1984, Thm. 3.3] along with Lemma 1. Let

$$c(T_L) = \gcd\{\text{length of } \alpha\colon \alpha \text{ is a cycle in } \Gamma(T_L)\}.$$

It is easy to see from (2.8) that there is a one-to-one correspondence between cycles of T_L and cycles of H_L with corresponding cycles having the same length. Thus $c(T_L) = c(H_L)$. It is well known (see Varga [1962, p. 49] or Berman and Plemmons [1979, p. 35]) that H_L is primitive exactly when $c(H_L) = 1$. Further, the regular splitting $A = (D - L) - U$ partitions the elements of A between $D - L$ and U so that $\Gamma(A)$ is simply the union of the graphs $\Gamma(D - L)$ and $\Gamma(U)$. Thus from Schneider [1984, Thm. 3.4], it follows that $c(T_L)$ is the greatest common divisor of the numbers of edges which lie in $\Gamma(U)$ for all cycles in $\Gamma(A)$. Thus H_L is primitive if and only this greatest common divisor is one. The theorem now follows from Lemma 1. □

As a result of Theorem 1 we can state the following sufficient condition for Gauss-Seidel convergence.

COROLLARY 1. (Rose [1984, Corollary 3]). *If* $\Gamma(A)$ *has a monotone decreasing (increasing) cycle then the forward (backward) Gauss-Seidel iteration scheme converges for each $x^{(0)}$.*

Proof. If α is a monotone decreasing (increasing) cycle in $\Gamma(A)$ then u_α (l_α) is 1. Alternatively, it also follows from Rose [1984] that $H_L(H_U)$ has a nonzero diagonal term and is thus primitive. □

Remark. This corollary and several corollaries in Schneider [1984, Corollaries 3.6 through 3.10] essentially have hypotheses which insure, at least for Gauss-Seidel splittings, that the iteration matrix has a nonzero entry on its diagonal. For example, if A has any symmetric zero structure (see Funderlic and Plemmons [1984]) then convergence of both forward and backward Gauss-Seidel follows. The existence of a nonzero diagonal term is sufficient but, as we show in § 3 by example, not necessary for convergence. However, this idea underlies our next two corollaries.

COROLLARY 2. *There is a permutation matrix P such that the forward (backward) Gauss-Seidel iteration scheme for P^TAP converges for each $x^{(0)}$.*

Proof. Let α be any cycle of the irreducible singular M-matrix matrix A. Relabel the vertices so that α is monotone decreasing (increasing). Convergence then follows by Corollary 1. □

The final corollary suggests a simple scheme for forcing Gauss-Seidel convergence by modifying the diagonal elements of the splitting with a scalar multiple of I. As we explain later, this iteration is closely related to the SOR method and an alternate proof of convergence can be obtained using Theorem 3.4 in Buoni, Neumann and Varga [1982]. The method will be stated for forward Gauss-Seidel. A similar statement can of course be made for backward Gauss-Seidel.

COROLLARY 3. *Let $\varepsilon > 0$. If*

$$A = D - L - U = (D - L + \varepsilon I) - (U + \varepsilon I)$$

and

$$T_\varepsilon = (D - L + \varepsilon I)^{-1}(U + \varepsilon I), \tag{2.10}$$

then T_ε is convergent for every $\varepsilon > 0$.

Proof. Since A is a singular irreducible M-matrix, it is easy to see that T_ε is a nonnegative matrix with spectral radius 1, and the elementary divisors associated with the eigenvalue 1 are linear. Further, since $(D - L + \varepsilon I)^{-1}$ and $(U + \varepsilon I)$ both have a positive diagonal, so does T_ε. Thus T_ε is convergent by Rose [1984, Prop. 2]. □

We note that the iteration matrix T_ε given by (2.10) reduces to the usual SOR iteration matrix

$$\mathscr{L}_\omega = (I - \omega L)^{-1}[(I - \omega)I + \omega U]$$

by the transformation $\varepsilon = 1/\omega - 1$. However, the use of (2.10) avoids scaling L and U. There exist algorithms for finding the cycles of a graph, as in Aho, Hopcroft and Ullman [1974]. Thus it may be feasible to apply Theorem 1 to determine a priori for A the convergence of backward or forward Gauss-Seidel iterations. Alternatively, one could find a cycle and permute the rows and columns of A by a permutation matrix P so that convergence is assured in view of Corollary 1. The concept of ordering a Q-matrix A to assure convergence is also discussed in Kaufman [1983] in the context of queueing networks.

The rate of convergence of an iteration matrix T is governed by the parameter

$$\gamma(T) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } T \text{ and } \lambda \neq 1\}$$

(see e.g., Berman and Plemmons [1979, Chap. 7]). The choice of the permutation matrix P in the above discussion to minimize $\gamma(T_L)$ or $\gamma(T_U)$ remains an open question. Some results for choosing the SOR parameter ω (or equivalently ε) in order to minimize $\gamma(T)$ have been given by Hadjidimos [1984] for certain A. The paper is concluded with some numerical examples which illustrate several of the ideas developed in this section concerning the convergence of Gauss-Seidel iterations.

In Buoni, Neumann and Varga [1982] there is an example, due to Hans Schneider, of a Q-matrix A for which (forward) Gauss-Seidel does not converge, but the Jacobi method, with the splitting $A = I - (L + U)$, does converge. This example answered in part a question in Neumann and Plemmons [1978]. Since we require zero column sums rather than zero row sums, the following matrix is the transpose of theirs. Specifically let

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

Then $(I - U)^{-1}L$ is the transpose of the matrix in Buoni, Neumann and Varga [1982, eq. (4.2), p. 194]. However, A has a monotone decreasing cycle, viz. (5, 4, 3, 2, 1, 5), and so (forward) Gauss-Seidel does converge for A by Corollary 1. In fact,

$$T_L = \begin{bmatrix} 0 & 0 & \frac{1}{2} & 0 & 1 \\ 0 & 0 & \frac{1}{2} & 0 & 1 \\ 0 & 0 & \frac{1}{2} & 0 & 1 \\ 0 & 0 & \frac{1}{4} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix}$$

and the eigenvalues are 0 and 1. Thus T_L is idempotent and (forward) Gauss-Seidel converges in one iteration.

It would be tempting to conjecture that for a Q-matrix either T_U or T_L will always converge. Unfortunately, this is false. For if

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

then

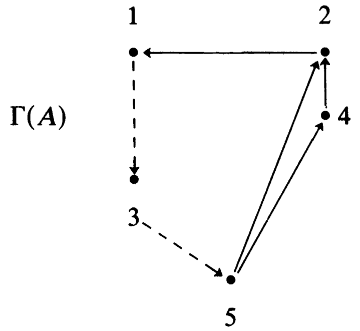
$$(I - L)^{-1}U = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \qquad (I - U)^{-1}L = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix},$$

are both cyclic and so neither converges. However, if rows and columns 1 and 3 or 2 and 4 of A are interchanged, then both forward and backward Gauss-Seidel converge for the resulting matrices.

As we have seen, several sufficient conditions for convergence have been stated in the literature by giving conditions under which the iteration matrix viz. (2.5), has a positive diagonal entry. This need not always be necessary for convergence. Let

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

with the graph



in which the edges from $\Gamma(L)$ are given by solid lines and those from $\Gamma(U)$ by broken lines. We have two cycles

$$\alpha=(1,3,5,4,2,1), \qquad u_\alpha=2,\ l_\alpha=3,$$

$$\beta=(1,3,5,2,1), \qquad u_\beta=2,\ l_\beta=2.$$

Thus $\gcd\{u_\alpha, u_\beta\}=2$ and $\gcd\{l_\alpha l_\beta\}=1$, so by Theorem 1, $T_L=(I-L)^{-1}U$ does not converge, while $T_U=(I-U)^{-1}L$ converges. In fact

$$T_L=\begin{bmatrix}0&0&1&0&0\\0&0&1&0&0\\0&0&0&0&1\\0&0&\frac{1}{2}&0&0\\0&0&1&0&0\end{bmatrix}, \qquad T_U=\begin{bmatrix}0&\frac{1}{2}&0&1&0\\1&0&0&0&0\\0&\frac{1}{2}&0&1&0\\0&\frac{1}{2}&0&0&0\\0&\frac{1}{2}&0&1&0\end{bmatrix},$$

with spectra $\sigma(T_L)=\{0,0,0,-1,1\}$ and $\sigma(T_U)=\{0,0,(1\pm i)/2,1\}$. Note in particular, that although T_U is convergent, it has all 0's on its diagonal, illustrating our point.

Acknowledgment. The authors would like to thank Professor Hans Schneider for helpful comments concerning his 1984 paper referenced herein.

REFERENCES

A. Aho, J. Hopcroft and J. Ullman [1974], *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA.

J. Barlow [1986], *On the smallest positive singular value of a singular M-matrix with applications to ergodic Markov chains*, this Journal, 7, pp. 414-424.

A. Berman and R. J. Plemmons [1979], *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York.

J. Buoni [1986], *Incomplete factorization of singular M-matrices*, this Journal, 7, pp. 193-198.

J. Buoni, M. Neumann and R. Varga [1982], *Theorems of Stein-Rosenberg type*: III, *the singular case*, Linear Algebra and Appl., 42, pp. 183-198.

P. Courtois and P. Semal [1985], *Block iterative algorithms for stochastic matrices*, preprint.

R. Funderlic and J. Mankin [1981], *Solution of homogeneous systems of linear equations arising from compartmental models*, SIAM J. Sci. Stat. Comput., 2, pp. 375-383.

R. Funderlic and C. D. Meyer, Jr. [1985], *Sensitivity of the stationary distribution vector of a Markov chain*, Linear Algebra Appl., to appear.

R. Funderlic and R. J. Plemmons [1984], *A combined direct-iterative method for certain M-matrix linear systems*, this Journal, 5, pp. 33-42.

G. H. Golub and C. D. Meyer, Jr. [1985], *Simultaneous computation of stationary probabilities with estimates of their sensitivities*, preprint.

A. HADJIDIMOS [1984], *Optimum iterative methods for the solution of singular linear systems*, TR 102, Dept. Mathematics, Univ. Ioannina, Greece.

W. HARROD AND R. J. PLEMMONS [1984], *Comparison of some direct methods for computing stationary distributions of Markov chains*, SIAM J. Sci. Stat. Comput., 5, pp. 453–469.

L. KAUFMAN [1983], *Matrix methods for queueing problems*, SIAM J. Sci. Stat. Comput., 4, pp. 525–552.

R. KOURY, D. MCALLISTER AND W. STEWART [1984], *Methods for computing stationary distributions of nearly completely decomposible Markov chains*, this Journal, 5, pp. 164–186.

M. NEUMANN AND R. J. PLEMMONS [1978], *Convergent nonnegative matrices and iterative methods for consistent linear systems*, Numer. Math., 31, pp. 173–186.

B. LUBACHEUSKI AND D. MITRA [1985], *A chaotic, asynchronous algorithm for computing the fixed point of a nonnegative matrix of unit spectral radius*, J. ACM, to appear.

D. ROSE [1984], *Convergent regular splittings for singular M-matrices*, this Journal, 5, pp. 133–144.

H. SCHNEIDER [1984], *Theorems on M-splittings of a singular M-matrix which depend on graph structure*, Linear Algebra Appl., 58, pp. 407–424.

G. STEWART, W. STEWART AND D. MCALLISTER [1985], *A two-stage iteration for solving nearly uncoupled Markov chains*, IEEE Trans. Software Engng., to appear.

R. VARGA [1962], *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

ON AN INVARIANT OF GRAPHS AND THE RELIABILITY POLYNOMIAL*

A. SATYANARAYANA† AND ZOHEL KHALIL‡

Abstract. A combinatorial invariant called the parity of a graph is introduced. An earlier concept of signed domination of a graph, relevant to computing certain reliability measures on an undirected graph, is a special case of this parity. A generalization of the signed domination theorem is presented. The results of this paper also provide some insight into the nature of the coefficient of the reliability polynomial.

Key words. network reliability, domination of a graph, reliability polynomial, factoring theorem

1. Introduction. A recent combinatorial invariant, called the domination of a graph, has played an important role in network reliability analysis [1], [2], [4], [5], [7]-[10]. Let $G=(V, E)$ be an undirected graph and $K \subseteq V$ be any subset of vertices such that $|K| \geqq 2$. A *K-tree* is a tree of G covering all vertices of K and whose pendant vertices are all in K. Alternatively, a K-tree is a minimal connected subgraph of G containing all vertices of K; it is minimal in the sense that deletion of any edge disconnects a pair of vertices in K. Any edge of G is *relevant* if it is in at least one K-tree of G. A *K-graph* is a graph in which all edges are relevant. Given that the edges of G may fail with known probabilities, the *K-terminal reliability* of G, denoted by $R_K(G)$, is the probability that G contains a K-tree in which all edges are working. A *formation U* is a set of K-trees of G whose union yields G. U is *odd* or *even* depending upon whether U contains *odd* or *even* number of trees respectively. The *signed domination* of G, with respect to some K, denoted by $d_K(G)$, is the number of odd minus the number of even formations of G. The absolute value, $|d_K(G)|$, has been used in [9] as a measure of complexity for computing $R_K(G)$. Another application of $d_K(G)$ is in the topological formula of $R_K(G)$ where the dominations of the K-subgraphs of G determine the coefficients of the terms in the formula [7], [8], [10].

In this paper, we introduce the concept of *parity* of a graph. Specifically, if S_i denotes the set of all subgraphs of G such that each subgraph in S_i has exactly i edges of G, then the i-parity of G, denoted by $\mathbf{P}_i(G)$, is equal to $\sum_{G_j \in S_i} d_k(G_j)$. Clearly, $d_K(G)=\mathbf{P}_{|E|}(G)$ and the concept of domination is a special case of the i-parity of G.

In [9], the following theorem has been proved.

THEOREM 1 (Satyanarayana and Chang). *Let G be a K-graph with respect to some K. If $G-e$ and G_e are the graphs obtained by deleting and contracting, respectively, an edge e in G, then $d_K(G)=d_K(G_e)-d_K(G-e)$.*

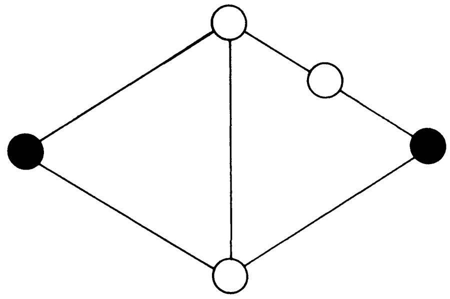
Set S_i can be partitioned into two subsets S_i^+ and S_i^- such that S_i contains all graphs of S_i having a specific edge e and S_i^- contains the remaining graphs of S_i. Let $\mathbf{P}_i^{e^+}(G)=\sum_{G_j \in S_i^+} d_K(G_j)$. We show that $\mathbf{P}_i^{e^+}(G)=\mathbf{P}_{i-1}(G_e)-\mathbf{P}_{i-1}(G-e)$. When $i=|E|$, Theorem 1 becomes a special case of this result. Further, an immediate consequence of this result is that $\mathbf{P}_i(G)=\mathbf{P}_i(G-e)+\mathbf{P}_{i-1}(G_e)-\mathbf{P}_{i-1}(G-e)$.

Suppose that the vertices of G do not fail and all the edges have equal working probabilities, say p. If the edge failure events are assumed to be statistically independent, then $R_K(G)$ can be expressed as a polynomial in p of at most degree $|E|$. An example graph and its reliability polynomial are shown in Fig. 1. We show that the coefficient

* Received by the editors November 7, 1984, and in final revised form June 6, 1985. This research was partially supported by the Natural Science and Engineering Research Council of Canada under grant A9095.

† Department of Computer Science, Stevens Institute of Technology, Hoboken, New Jersey 07030.

‡ Department of Mathematics, Concordia University, Montreal, Quebec, Canada H3G 1M8.



$R_K(G)=p^2+2p^3-4p^5+2p^6$

FIG. 1. *An example graph and its reliability polynomial. Darkened vertices are K-vertices.*

of the ith degree term in the polynomial is equal to $\mathbf{P}_i(G)$, thus providing some insight on the composition of the coefficients in the polynomial.

2. Preliminaries. Graph theoretic terminology used here generally follows Bondy and Murty [3]. We are only concerned with undirected graphs.

Consider a graph $G=(V,E)$ with vertex set $V=\{v_1, v_2, \cdots, v_n\}$ and edge set $E=\{e_1, e_2, \cdots, e_b\}$. Let K be a specified subset of V with $|K|\geqq 2$. A *K-tree* is a tree of G covering all vertices of K such that its pendant vertices are in K. For the two special cases of $|K|=2$ and $|K|=|V|$, a K-tree is a simple path and a spanning tree respectively. A graph G, with respect to some K, is a *K-graph* if every edge of G is in some K-tree. A *formation* U of a graph G is a set of K-trees of G whose union yields the graph G. U is *odd* (*even*) depending upon whether the number of trees in U is odd (even) respectively. The *signed domination*, denoted by $d_K(G)$, of a graph G is the number of odd minus the number of even formations of G. Let $S_i=\{G_1, G_2, \cdots, G_m\}$ be the set of all subgraphs of G such that each subgraph $G_i \in S_i$ has exactly i edges. The parity of G, denoted by $\mathbf{P}_i(G)$, is the sum of the signed dominations of all subgraphs in S_i. In other words, $\mathbf{P}_i(G)=\sum_{G_j\in S_i} d_K(G_j)$.

Relative to any specific edge e in G, the set S_i can be partioned into two subsets S_i^+ and S_i^-, such that S_i^+ contains all graphs of S_i having e and S_i^- contains all those without e. Let

$$\mathbf{P}_i^{e^+}(G)=\sum_{G_j\in S_i^+} d_K(G_j) \quad \text{and} \quad \mathbf{P}_i^{e^-}(G)=\sum_{G_j\in S_i^-} d_K(G_j).$$

Note that $\mathbf{P}_i(G)=\mathbf{P}_i^{e^+}(G)+\mathbf{P}_i^{e^-}(G)$.

If $i=|E|$, then the only element in S_i is G itself. Thus $\mathbf{P}_i(G)=d_K(G)$ and also $\mathbf{P}_i^{e^+}(G)=d_K(G)$.

The following proposition is relevant to our discussion; this has been proved in the context of signed domination in [9].

PROPOSITION 1. *For any graph G with a specified K, $d_K(G)\neq 0$ if and only if G is a K-graph.*

In view of Proposition 1, without loss of generality, the set S_i can be redefined as the one constituting all K-subgraphs of G with i edges.

If e is an edge with end vertices u and v in G, then $G-e$ is a subgraph of G obtained by deleting e from G. Deletion of an edge does not imply deletion of its end vertices. In G, contracting edge e involves deleting e and merging vertices u and v into a supervertex, such that every edge that was incident on either u or v or both is incident on the supervertex. The graph obtained from G by contracting e is denoted by G_e.

3. A theorem on parity. We first show that $\mathbf{P}_i^{e^+} = \mathbf{P}_{i-1}(G_e) - \mathbf{P}_{i-1}(G-e)$. Since $\mathbf{P}_{|E|}^{e^+}(G) = d_K(G)$, $\mathbf{P}_{|E|-1}(G_e) = d_K(G_e)$ and $\mathbf{P}_{|E|-1}(G-e) = d_K(G-e)$, this result embodies an important generalization of Theorem 1 [9]. Furthermore, a direct consequence of this result, as will be shown in Corollary 1, is a recursive formula for computing $\mathbf{P}_i(G)$.

THEOREM 2. *Let G be graph with a specified K. If $G-e$ and G_e are the graphs obtained by deleting and contracting, respectively an edge e in G, then*

$$\mathbf{P}_i^{e^+}(G) = \mathbf{P}_{i-1}(G_e) - \mathbf{P}_{i-1}(G-e).$$

Proof. Suppose $S_i^+ = \{G_1, G_2, \cdots, G_m\}$, is the set of all K subgraphs of G containing e. From any $G_j \in S_i^+$, we can obtain two graphs $G_{j,e}$ and $G_j - e$ by contracting and deleting e respectively. Define $X_{i-1}^c = \{G_{1,e}, G_{2,e}, \cdots, G_{m,e}\}$ the set of graphs obtained from S_i^+ by contracting e. Similarly, $X_{i-1}^D = \{(G_1 - e), (G_2 - e), \cdots, (G_m - e)\}$ is the set of graphs obtained by deleting e. Every graph in X_{i-1}^c is a subgraph of G_e. However, it is not necessary that all K-subgraphs of G_e are in X_{i-1}^c. Therefore, let $Y_{i-1} = \{H_1, H_2, \cdots, H_h\}$ constitute all the K-subgraphs having $i-1$ edges of G_e that are not in X_{i-1}^c. By definition

$$\mathbf{P}_{i-1}(G_e) = \sum_{G_{j,e} \in X_{i-1}^c} d_K(G_{j,e}) + \sum_{H_j \in Y_{i-1}} d_K(H_j). \tag{1}$$

Applying Theorem 1 on each of the graphs in S_i^+, we get

$$\mathbf{P}_i^{e^+}(G) = \sum_{G_{j,e} \in X_{i-1}^c} d_K(G_{j,e}) - \sum_{(G_j - e) \in X_{i-1}^D} d_K(G_j - e). \tag{2}$$

From (1) and (2),

$$\mathbf{P}_i^{e^+}(G) = \mathbf{P}_{i-1}(G_e) - \left(\sum_{(G_j - e) \in X_{i-1}^D} d_K(G_j - e) + \sum_{H_j \in Y_{i-1}} d_K(H_j) \right). \tag{3}$$

Now we need only to show that

$$\sum_{G_j - e \in X_{i-1}^D} d_K(G_j - e) + \sum_{H_j \in Y_{i-1}} d_K(H_j) = \mathbf{P}_{i-1}(G-e). \tag{4}$$

Suppose v_1 and v_2 are the end vertices of edge e. Every K-subgraph of $G-e$ having $i-1$ edges and containing both v_1 and v_2 is in X_{i-1}^D. Every K-subgraph of $G-e$ having $i-1$ edges and not containing at least one of v_1 or v_2 is in Y_{i-1}. Thus the LHS of (4) contains all the K-subgraphs of $G-e$ having $i-1$ edges. However, it is possible that the LHS of (4) may contain some non$K$-subgraphs, but it can be easily shown that every K-subgraph in the LHS is also a K-subgraph of $G-e$. Hence, we have (4) and substituting (4) in (3), we have the theorem. □

COROLLARY 1. *Let G be any graph with a specified K and e be any edge in G. Then*

$$\mathbf{P}_i(G) = \mathbf{P}_i(G-e) + \mathbf{P}_{i-1}(G_e) - \mathbf{P}_{i-1}(G-e).$$

Proof. This follows from the facts, $\mathbf{P}_i(G) = \mathbf{P}_i^{e^+}(G) + \mathbf{P}_i^{e^-}(G)$, $\mathbf{P}_i^{e^-}(G-e) = \mathbf{P}_i(G-e)$, and by Theorem 2. □

4. Coefficients of the reliability polynomial. Suppose $\{t_1, t_2, \cdots, t_m\}$ is the set of K-trees of G with respect to a given K. Then by the inclusion-exclusion principle we have,

$$R_K(G) = \sum_i \Pr\{t_i\} - \sum_{i<j} \Pr\{t_i \cap t_j\} + \sum_{i<j<k} \Pr\{t_i \cap t_j \cap t_k\} - \cdots. \tag{5}$$

In (5), an event of the form $\{t_i \cap t_j \cap \cdots \cap t_k\}$ is an event that all the elements in the K-subgraph $(t_i \cup t_j \cup \cdots \cup t_k)$ of G are operative.

The number of terms in (5) is $2^m - 1$. Indeed every term in (5) corresponds to some formation of a K-subgraph of G. Specifically, define an *i-formation* $U^{(i)}$ as a set of K-trees of G whose union yields any K-subgraph having exactly i edges of G. $U^{(i)}$ is odd (even) depending upon whether $U^{(i)}$ contains odd (even) number of trees. It can be easily seen that there exists a one-to-one correspondence between the terms in (5) and all possible formations of K-subgraphs of G.

Any $U^{(i)}$ yields exactly one K-subgraph of G. However, it is possible that two i-formations $U_1^{(i)}$ and $U_2^{(i)}$ might either yield the same K-subgraph or two different K-subgraphs having i edges of G. Let θ be the set of 2^{m-1} possible i-formations. Partition $\theta = \{\theta_1, \theta_2, \cdots, \theta_h\}$ such that any i-formation $U^{(i)} \in \theta_j$, $j = 1, \cdots, h$, yields the K-subgraph G_j of G. Thus the set $\{G_1, G_2, \cdots, G_h\}$ constitute all possible K-subgraphs of G and

$$R_K(G) = \sum_{j=1}^{h} (d_K(G_j) \cdot \Pr(G_j)), \tag{6}$$

where $\Pr(G_j)$ is the probability that all edges in G_j are working [7]. Note that the number of terms in (6) is far less than those in (5) and the coefficient associated with any term in (6) is the signed domination of the corresponding K-subgraph.

In the case of directed graphs, it has been shown that, if G_j is acyclic then $d_K(G_j) = (-1)^{E_j - V_j + 1}$, where E_j and V_j are respectively the number of edges and the number of vertices in the K-subgraph G_j; for all cyclic graphs $d_K(G_j) = 0$ [8], [10]. In our case, where G is undirected, $d_K(G) \neq 0$ and is always an integer not necessarily equal to ±1.

If all the edges of G have equal reliabilities, say p, and the edge failures are statistically independent, then (6) reduces to:

$$R_K(G) = \sum \mathbf{P}_i(G) p^i. \tag{7}$$

The right-hand side of (7) is a polynomial in p containing at most $|E|$ terms. The coefficients of the ith degree term in (7) is given by the i-parity of G.

Corollary 1 provides a topological interpretation and provides some insight into the composition of the coefficients of (7). Furthermore, using Corollary 1 and (7) we have

$$\begin{aligned} R_K(G) &= \sum (\mathbf{P}_i(G-e))p^i + p \sum (\mathbf{P}_{i-1}(G_e))p^{i-1} - p \sum (\mathbf{P}_{i-1}(G-e))p^{i-1} \\ &= pR_K(G_e) + (1-p)R_K(G-e), \end{aligned}$$

which is the well-known factoring theorem [6], [9].

REFERENCES

[1] A. Agrawal and R. E. Barlow, *A survey of network reliability and domination theory*, Oper. Res., 32 (1984), pp. 478–492.

[2] R. E. Barlow, *Set theoretic signed domination for coherent structures*, Technical Report Number ORC 82-1, Operations Research Center, Univ. California, Berkeley, 1982.

[3] J. A. Bondy and U. S. R. Murty, *Graph Theory with Applications*, North-Holland, New York, 1976.

[4] A. B. Huseby, *A generalized optimal factoring algorithm for computing exact system reliability*, Proc. S. R. E. Symposium, Arboga, 1983.

[5] ——— *A unified theory of domination and signed domination with application to exact reliability computations*, Statistical Research Report, Institute of Mathematics, University of Oslo, Norway, 1984.

[6] F. MOSKOWITZ, *The analysis of redundancy networks*, AIEE Trans. Commun. Electron., 39 (1958), pp. 627-632.

[7] A. SATYANARAYANA, *Multi-terminal network reliability*, Technical Report Number ORC 80-6, Operations Research Center, Univ. California, Berkeley, 1980.

[8] ——— *A unified formula for analysis of some network reliability problems*, IEEE Trans. Reliability, R-31 (1982) 23-32.

[9] A. SATYANARAYANA AND M. K. CHANG, *Network reliability and the factoring theorem*, Networks, 13 (1983), pp. 107-120.

[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.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

009

WEIGHTED AVERAGES OF RADON TRANSFORMS ON Z_2^k*

J. A. MORRISON†

Abstract. Weighted averages of Radon transforms on the group of binary k-tuples under modulo 2 addition, which arise in applied statistics, are investigated. The inversion formula is derived by means of (discrete) Fourier transforms, and also by means of an expansion in terms of Krawtchouk polynomials. Alternate explicit representations are obtained for the coefficients in the inversion formula in several particular cases. Moreover, when the Radon transform is over the subset of Hamming distance 2, the asymptotic behavior of these coefficients is investigated when k is large (and not the square of an integer).

Key words. Gelfand pairs, Krawtchouk polynomials, spherical functions

AMS(MOS) subject classifications. 42A76, 33A65, 33A45, 20C99

1. Introduction. Recently, Diaconis and Graham [1] have investigated the Radon transform on Z_2^k, the group of binary k-tuples under modulo 2 addition. Discrete Radon transforms of this type arise in applied statistics, and many examples have been discussed by Diaconis [2], [3]. For a function $f: Z_2^k \to R$, and a nonempty subset $S \subset Z_2^k$, the Radon transform is defined as

$$\bar{f}(y) = \sum_{x \in S+y} f(x). \tag{1.1}$$

The Fourier transform of the indicator function of the set S is

$$\hat{\chi}_S(y) = \sum_{x \in S} (-1)^{x \cdot y}. \tag{1.2}$$

Diaconis and Graham showed that the transform $f \to \bar{f}$ is one-to-one if and only if $\hat{\chi}_S(y) \neq 0$ for any $y \in Z_2^k$, and they derived the inverse with the help of Fourier transforms. Let $H(x, y)$ denote Hamming distance, the number of coordinates where x and y disagree, and let $H(x) = H(x, 0)$. Then, as examples, they considered $S = S_r = \{x \in Z_2^k\colon H(x) = r\}$ and $S = S_r^+ = \{x \in Z_2^k\colon H(x) \leqq r\}$, i.e., averages of f over all points of distance equal to r, and averages over all points of distance less than or equal to r. In particular, they obtained simple inversion formulas when k is odd and $S = S_1$, and when k is even and $S = S_1^+$.

In this paper we consider the more general transform, involving weighted averages of the Radon transform on Z_2^k for the sets S_r,

$$\bar{f}(y) = \sum_{r=0}^{k} \alpha_r \sum_{H(x,y)=r} f(x) \tag{1.3}$$

where α_r, $r = 0, \cdots, k$ are real constants, not all of which are zero. In § 2, following Diaconis and Graham [1], we derive the inversion formula by means of Fourier transforms. The coefficients in the formula

$$f(0) = \sum_{w=0}^{k} c_w \sum_{H(z)=w} \bar{f}(z) \tag{1.4}$$

are given explicitly in terms of Krawtchouk polynomials [4, p. 130]. In § 3 alternate explicit representations are obtained for these coefficients in several particular cases, including $S = S_1$, S_1^+, S_2, S_3 and S_4 in (1.1). The results are shown to reduce to those

* Received by the editors April 29, 1985.

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

found by Diaconis and Graham [1] when $S=S_1$ and $S=S_1^+$. In § 4 we consider the case $S=S_2$ in (1.1), for which $c_{2l+1}=0$ in (1.4), and investigate the asymptotic behavior of the coefficients c_{2l} when k is large (and not the square of an integer). The asymptotic approximations are derived separately for (i) $l=O(1)$, (ii) $k-2l=O(1)$, and (iii) $2l=k\sin^2\theta$ where $0<\theta<\pi/2$, and $\sqrt{k}\,\theta\gg 1$ and $\sqrt{k}(\pi/2-\theta)\gg 1$. It is shown that the approximation corresponding to (iii) matches with the other two approximations for suitably small θ, and $(\pi/2-\theta)$, respectively. The approximations indicate that, for $k\gg 1$, the coefficient with the largest magnitude is either c_0 or $c_{2\lfloor k/2\rfloor}$. Finally, in § 5, we give an alternative derivation of the inversion formula, by means of an expansion in terms of Krawtchouk polynomials.

2. Inversion formula. We follow Diaconis and Graham [1] and write the transform (1.3) as a convolution

$$\bar{f}(y)=\sum_{r=0}^{k}\alpha_r\sum_{H(z)=r}f(y-z)=f*\chi(y), \tag{2.1}$$

where

$$\chi(z)=\sum_{r=0}^{k}\alpha_r\chi_r(z), \tag{2.2}$$

and $\chi_r(z)$ is the indicator function of the set $S_r=\{z\in Z_2^k\colon H(z)=r\}$. For a function $F\colon Z_2^k\to R$ we have the Fourier transform, and inverse,

$$\hat{F}(y)=\sum_x(-1)^{x\cdot y}F(x),\qquad F(x)=\frac{1}{2^k}\sum_y(-1)^{x\cdot y}\hat{F}(y). \tag{2.3}$$

Hence, from (2.1)-(2.3), we obtain

$$\hat{\bar{f}}(z)=\hat{f}(z)\hat{\chi}(z),\qquad \hat{\chi}(z)=\sum_{r=0}^{k}\alpha_r\hat{\chi}_r(z). \tag{2.4}$$

Also, as shown in [1],

$$\hat{\chi}_r(z)=p_r^k(H(z)), \tag{2.5}$$

where, for ν, $r=0,\cdots,k$, $p_r^k(\nu)$ is the Krawtchouk polynomial [4, p. 130]

$$p_r^k(\nu)=\sum_{j=0}^{r}(-1)^j\binom{\nu}{j}\binom{k-\nu}{r-j}. \tag{2.6}$$

We remark that the kernel $\chi(z)$ is a spherical function, since the indicator functions $\chi_r(z)$ are invariant under the natural action of the permutation group Π_k on Z_2^k. The group $\Pi_k\times Z_2^k$, where $(\pi,z)x=\pi x+z$, and its subgroup Π_k is an example of a Gelfand pair (see Letac [6] for this and other examples).

We will assume that

$$P(\nu)=\sum_{r=0}^{k}\alpha_r p_r^k(\nu)\neq 0,\qquad \nu=0,\cdots,k, \tag{2.7}$$

so that $\hat{\chi}(z)\neq 0$. Then, from (2.3) and (2.4), it follows that

$$f(x)=\frac{1}{2^k}\sum_y(-1)^{x\cdot y}\frac{\hat{\bar{f}}(y)}{\hat{\chi}(y)}=\frac{1}{2^k}\sum_y\sum_z(-1)^{y\cdot z}\frac{\bar{f}(z-x)}{\hat{\chi}(y)}. \tag{2.8}$$

It will suffice to determine $f(0)$, since an expression for $f(x)$ may then be obtained by shifting everything by x. But, from (2.4), (2.5), (2.7) and (2.8), we obtain

$$f(0)=\frac{1}{2^k}\sum_{\nu=0}^{k}\sum_{H(y)=\nu}\sum_{w=0}^{k}\sum_{H(z)=w}(-1)^{y\cdot z}\frac{\bar{f}(z)}{P(\nu)}. \tag{2.9}$$

As in [1], we define

$$\bar{g}(w)=\sum_{H(z)=w}\bar{f}(z). \tag{2.10}$$

Hence

$$f(0)=\sum_{w=0}^{k}c_w\bar{g}(w), \tag{2.11}$$

where, counting the number of times for which $y\cdot z=u$ when $H(y)=\nu$, $H(z)=w$ and z is fixed,

$$c_w=\frac{1}{2^k}\sum_{\nu=0}^{k}\sum_{u=0}^{\nu}\frac{(-1)^u}{P(\nu)}\binom{w}{u}\binom{k-w}{\nu-u}=\frac{1}{2^k}\sum_{\nu=0}^{k}\frac{p_\nu^k(w)}{P(\nu)}. \tag{2.12}$$

The above expression for the coefficients c_w in the inversion formula $f(0)$ is explicit, in view of (2.6) and (2.7). However, Diaconis and Graham [1] obtained simple formulas for c_w corresponding to the Radon transform (1.1) when $S=S_1=\{x\in Z_2^k\colon H(x)=1\}$ and k is odd, and when $S=S_1^+=\{x\in Z_2^k\colon H(x)\leqq 1\}$ and k is even. They did this by defining

$$g(u)=\sum_{H(x)=u}f(x), \tag{2.13}$$

and considering the transform $g\to\bar{g}$, where $\bar{g}$ is defined by (2.10). In the two cases mentioned they were able to solve the recurrence relations to obtain the coefficients of $\bar{g}(w)$, $w=0,\cdots,k$, in the expression for $g(0)=f(0)$. We will show how to obtain their results directly from (2.12), which may be written in the form

$$c_w=\frac{1}{2^k}\sum_{u=0}^{w}\sum_{s=0}^{k-w}\frac{(-1)^u}{P(u+s)}\binom{w}{u}\binom{k-w}{s}. \tag{2.14}$$

3. Particular cases. We first consider the case $S=S_1$ in (1.1), corresponding to $\alpha_1=1$, and $\alpha_r=0$ for $r\neq 1$, in (1.3). Then, from (2.6) and (2.7), we have $P(\nu)=k-2\nu$ and $k=2m+1$ is odd. We make use of the representation

$$\frac{1}{P(\nu)}=\frac{1}{(2m+1-2\nu)}=\frac{1}{2\pi i}\int_{-\pi}^{\pi}t\,e^{i(2m+1-2\nu)t}\,dt. \tag{3.1}$$

But,

$$\sum_{u=0}^{w}\sum_{s=0}^{2m+1-w}(-1)^u\binom{w}{u}\binom{2m+1-w}{s}e^{i(2m+1-2u-2s)t}=(e^{it}-e^{-it})^w(e^{it}+e^{-it})^{2m+1-w}. \tag{3.2}$$

Hence, from (2.14), (3.1) and (3.2), we obtain

$$c_w=\frac{i^{w-1}}{2\pi}\int_{-\pi}^{\pi}t\sin^w t\cos^{2m+1-w}t\,dt. \tag{3.3}$$

It follows that $c_{2l}=0$ and

$$\begin{aligned} c_{2l+1}&=\frac{(-1)^l}{\pi}\int_0^{\pi} t\sin^{2l+1}t\cos^{2m-2l}t\,dt\\ &=(-1)^l\int_0^{\pi/2}\sin^{2l+1}t\cos^{2m-2l}t\,dt, \end{aligned} \tag{3.4}$$

where we have changed t to $(\pi-t)$ in the interval $(\pi/2,\pi)$. Therefore [5, p. 8]

$$c_{2l+1}=\frac{(-1)^l l!\Gamma(m-l+\frac{1}{2})}{2\Gamma(m+\frac{3}{2})}=\frac{(-1)^l l!\Gamma(m-l+\frac{1}{2})}{(2m+1)\Gamma(m+\frac{1}{2})}, \tag{3.5}$$

as found in [1].

We now consider the case $S=S_1^+$ in (1.1), corresponding to $\alpha_0=\alpha_1=1$, and $\alpha_r=0$ for $r\geqq 2$, in (1.3). Then, from (2.6) and (2.7), we have $P(\nu)=k+1-2\nu$ and $k=2m$ is even, so that $P(\nu)$ is given by (3.1). But,

$$\sum_{u=0}^{w}\sum_{s=0}^{2m-w}(-1)^u\binom{w}{u}\binom{2m-w}{s}e^{i(2m+1-2u-2s)t}=e^{it}(e^{it}-e^{-it})^w(e^{it}+e^{-it})^{2m-w}. \tag{3.6}$$

Hence, from (2.14), (3.1) and (3.6), we obtain

$$c_w=\frac{i^{w-1}}{2\pi}\int_{-\pi}^{\pi} t\,e^{it}\sin^w t\cos^{2m-w}t\,dt. \tag{3.7}$$

It follows that

$$c_{2l}=c_{2l+1}=\frac{(-1)^l}{\pi}\int_0^{\pi} t\sin^{2l+1}t\cos^{2m-2l}t\,dt. \tag{3.8}$$

Moreover, c_{2l+1} is given by (3.5), in view of (3.4). This agrees with the result obtained in [1].

Next we consider the case $S=S_2$ in (1.1), corresponding to $\alpha_2=1$, and $\alpha_r=0$ for $r\neq 2$, in (1.3). Then, from (2.6) and (2.7), we have $P(\nu)=\frac{1}{2}[(k-2\nu)^2-k]$, and $k\neq m^2$, where m is an integer. Hence,

$$\frac{1}{P(\nu)}=\frac{1}{\sqrt{k}}\left[\frac{1}{(k-\sqrt{k}-2\nu)}-\frac{1}{(k+\sqrt{k}-2\nu)}\right]. \tag{3.9}$$

We define

$$d_w^k(\lambda)=\frac{1}{2^k}\sum_{u=0}^{w}\sum_{s=0}^{k-w}\frac{(-1)^u}{(k+\lambda-2u-2s)}\binom{w}{u}\binom{k-w}{s}, \tag{3.10}$$

for $\lambda\neq\pm m$. Then, from (2.14), we obtain

$$c_w=\frac{1}{\sqrt{k}}[d_w^k(-\sqrt{k})-d_w^k(\sqrt{k})]. \tag{3.11}$$

But,

$$\int_{-\pi}^{\pi}e^{i(k+\lambda-2\nu)t}\,dt=\frac{2(-1)^k\sin(\lambda\pi)}{(k+\lambda-2\nu)}. \tag{3.12}$$

Hence, from (3.10), if $\lambda \neq \pm m$, it follows that

$$(3.13)\qquad d^k_w(\lambda)=\frac{(-1)^k i^w}{2\sin(\lambda\pi)}\int_{-\pi}^{\pi} e^{i\lambda t}\sin^w t\cos^{k-w} t\,dt.$$

Therefore

$$(3.14)\qquad d^k_{2l}(\lambda)=\frac{(-1)^{k+l}}{\sin(\lambda\pi)}\int_0^{\pi}\cos(\lambda t)\sin^{2l} t\cos^{k-2l} t\,dt,$$

and

$$(3.15)\qquad d^k_{2l+1}(\lambda)=\frac{(-1)^{k+l+1}}{\sin(\lambda\pi)}\int_0^{\pi}\sin(\lambda t)\sin^{2l+1} t\cos^{k-2l-1} t\,dt.$$

Hence, from (3.11), we obtain $c_{2l+1}=0$ and

$$(3.16)\qquad c_{2l}=\frac{2(-1)^{k+l+1}}{\sqrt{k}\sin(\sqrt{k}\,\pi)}\int_0^{\pi}\cos(\sqrt{k}\,t)\sin^{2l} t\cos^{k-2l} t\,dt.$$

Since

$$\frac{d}{dt}\{\cos^{k-1} t[\sin(\sqrt{k}\,t)\cos t-\sqrt{k}\cos(\sqrt{k}\,t)\sin t]\}$$
$$=\sqrt{k}\,(k-1)\cos(\sqrt{k}\,t)\sin^2 t\cos^{k-2} t,$$

it follows that $c_2=1/\binom{k}{2}$, as it should.

We may obtain an alternate representation for $d^k_w(\lambda)$ as follows. If we make use of (3.12) in (3.10), and sum only on u, we obtain

$$(3.17)\qquad d^k_w(\lambda)=\frac{(-1)^k i^w}{2^{k-w+1}\sin(\lambda\pi)}\sum_{s=0}^{k-w}\binom{k-w}{s}\int_{-\pi}^{\pi}\sin^w t\,e^{i(k+\lambda-w-2s)t}\,dt.$$

But [5, p. 8],

$$(3.18)\qquad \int_0^{\pi}\sin^w t\,e^{i\beta t}\,dt=\frac{\pi w!\exp(i\beta\pi/2)}{2^w\Gamma(1+w/2+\beta/2)\Gamma(1+w/2-\beta/2)},$$

and [5, p. 2]

$$(3.19)\qquad \Gamma(z)\Gamma(1-z)=\pi\operatorname{cosec}(\pi z).$$

It is found, from (3.17)-(3.19), that

$$(3.20)\qquad d^k_w(\lambda)=-\frac{w!}{2^{k+1}}\sum_{s=0}^{k-w}\binom{k-w}{s}\frac{1}{(s-k/2-\lambda/2)_{w+1}},$$

where

$$(3.21)\qquad (a)_{w+1}=a(a+1)\cdots(a+w),\qquad w=0,1,\cdots.$$

In terms of the hypergeometric function [5, p. 39],

$$(3.22)\qquad d^k_w(\lambda)=\frac{-w!}{2^{k+1}(-k/2-\lambda/2)_{w+1}}F\left(-k+w,-\frac{k}{2}-\frac{\lambda}{2};1+w-\frac{k}{2}-\frac{\lambda}{2};-1\right).$$

It is clear that, in general, we may obtain an alternate representation for the coefficients c_w in the inversion formula (2.11), by expanding the reciprocal of $P(\nu)$ in (2.14) in partial fractions. We proceed to give the results for some other particular cases.

(i) $S=S_3$. This corresponds to $\alpha_3=1$, and $\alpha_r=0$ for $r\neq 3$, in (1.3). Then, from (2.6) and (2.7), we obtain

$$P(\nu)=\tfrac{1}{6}(k-2\nu)[(k-2\nu)^2-3k+2], \tag{3.23}$$

and $k=2m+1$ is odd, with $3k-2=6m+1\neq n^2$, where n is an integer. It is found that $c_{2l}=0$ and

$$c_{2l+1}=\frac{6}{(6m+1)}\left[d_{2l+1}^{2m+1}(\sqrt{6m+1})-\frac{(-1)^l l!\Gamma(m-l+\frac{1}{2})}{(2m+1)\Gamma(m+\frac{1}{2})}\right]. \tag{3.24}$$

(ii) $S=S_4$. This corresponds to $\alpha_4=1$, and $\alpha_r=0$ for $r\neq 4$, in (1.3). Then, from (2.6) and (2.7), we obtain

$$P(\nu)=\tfrac{1}{24}[(k-2\nu)^4-2(3k-4)(k-2\nu)^2+3k(k-2)]. \tag{3.25}$$

We let

$$\kappa=[(3k-4)+(6k^2-18k+16)^{1/2}]^{1/2}, \qquad \sigma=[(3k-4)-(6k^2-18k+16)^{1/2}]^{1/2}, \tag{3.26}$$

and note that κ and σ are positive, since $k\geqq 4$. Then

$$P(\nu)=\tfrac{1}{24}[(k-2\nu)^2-\kappa^2][(k-2\nu)^2-\sigma^2]. \tag{3.27}$$

If κ and σ are not integers, it is found that $c_{2l+1}=0$ and

$$c_{2l}=\frac{24}{(\kappa^2-\sigma^2)}\left[\frac{1}{\sigma}d_{2l}^k(\sigma)-\frac{1}{\kappa}d_{2l}^k(\kappa)\right]. \tag{3.28}$$

(iii) $\alpha_0=\alpha$, $\alpha_1=\beta$, $\alpha_2=1$, $\alpha_r=0$, $r\geqq 3$. In this case, from (2.6) and (2.7), we obtain

$$P(\nu)=\tfrac{1}{2}[(k+\beta-2\nu)^2-\mu^2], \qquad \mu=(k+\beta^2-2\alpha)^{1/2}. \tag{3.29}$$

If $\mu\neq 0$, then

$$c_w=\frac{1}{\mu}[d_w^k(\beta-\mu)-d_w^k(\beta+\mu)]. \tag{3.30}$$

Expressions for $d_w^k(\lambda)$ are given by (3.13) and by (3.20). However, the integral in (3.13) is zero if $\lambda=\pm(k+p)$, where p is a positive integer. The integral is also zero if $\lambda=k-2n+1$, $n=1,\cdots,k$. In these cases the limiting value must be taken in (3.13), since $\sin(\lambda\pi)=0$ also. On the other hand, the expression for $d_w^k(\lambda)$ in (3.20) is valid in these cases. The integral in (3.13) is generally not zero if $\lambda=2\nu-k$, $\nu=0,\cdots,k$. However, by assumption $(k+\beta-2\nu)^2\neq\mu^2$, i.e. $\beta\pm\mu\neq 2\nu-k$, $\nu=0,\cdots,k$. The result for $\mu=0$ may be obtained from the limit $\mu\to 0$.

4. $S=S_2$ and k large. We here consider the Radon transform over the subset of Hamming distance 2, i.e. $S=S_2$, and investigate the asymptotic behavior of the coefficients in the inversion formula (2.11), when k is large (and not the square of an integer). We have shown that $c_{2l+1}=0$, and that c_{2l} is given by (3.16) for $l=0,\cdots,\lfloor k/2\rfloor$. The asymptotic approximations of c_{2l} will be derived separately for (i) $l=O(1)$, (ii) $k-2l=O(1)$, and (iii) $2l=k\sin^2\theta$ where $0<\theta<\pi/2$, and $\sqrt{k}\,\theta\gg 1$ and $\sqrt{k}(\pi/2-\theta)\gg 1$. It will also be shown that the approximation corresponding to (iii) matches with the other two approximations for suitably small θ, and $(\pi/2-\theta)$, respectively.

(i) $l=O(1)$. In this case the main contributions to the integral in (3.16) occur near the endpoints, and we write

$$\begin{aligned}&\int_0^{\pi} \cos(\sqrt{k}\,t)\sin^{2l} t\cos^{k-2l} t\,dt\\ &\quad=\int_0^{\pi/2}\{\cos(\sqrt{k}\,t)+(-1)^k\cos[\sqrt{k}(\pi-t)]\}\sin^{2l}t\cos^{k-2l}t\,dt.\end{aligned}\tag{4.1}$$

But, for $\tau=O(1)$ and $l=O(1)$,

$$\sin^{2l}(\tau/\sqrt{k})\sim(\tau^2/k)^l,\qquad \cos^{k-2l}(\tau/\sqrt{k})\sim e^{-\tau^2/2}\quad\text{for } k\gg 1.\tag{4.2}$$

We define

$$A_r=\int_0^{\infty}\tau^r e^{-\tau^2/2}\cos\tau\,d\tau,\qquad B_r=\int_0^{\infty}\tau^r e^{-\tau^2/2}\sin\tau\,d\tau.\tag{4.3}$$

It follows, from (3.16) and (4.1)-(4.3), that

$$c_{2l}\sim\frac{2(-1)^{l+1}}{k^{l+1}}\left\{\frac{[\cos(\sqrt{k}\,\pi)+(-1)^k]}{\sin(\sqrt{k}\,\pi)}A_{2l}+B_{2l}\right\},\qquad l=O(1).\tag{4.4}$$

It may be shown, by integrating by parts, that $A_2=0$ and $B_2=1$, so that $c_2\sim 2/k^2$, as it should.

(ii) $k-2l=O(1)$. In this case the main contribution to the integral in (3.16) occurs near the midpoint, and with $k-2l=m$ we write

$$\begin{aligned}&\int_0^{\pi}\cos(\sqrt{k}\,t)\sin^{k-m}t\cos^m t\,dt\\ &\quad=\int_0^{\pi/2}\{\cos\{\sqrt{k}(\pi/2-t)]+(-1)^m\cos[\sqrt{k}(\pi/2+t)]\}\cos^{k-m}t\sin^m t\,dt.\end{aligned}\tag{4.5}$$

If we let $\sqrt{k}\,t=\tau$ and proceed as before, we find from (3.16), (4.3) and (4.5), with $m=2q+k-2\lfloor k/2\rfloor$, that

$$\begin{aligned}c_{2(\lfloor k/2\rfloor-q)}\sim\frac{2(-1)^{\lfloor k/2\rfloor-q+1}}{k^{q+1}\sin(\sqrt{k}\,\pi)}&\Big\{[1+(-1)^k]\cos(\sqrt{k}\,\pi/2)A_{2q}\\ &\quad-\frac{[1-(-1)^k]}{\sqrt{k}}\sin(\sqrt{k}\,\pi/2)B_{2q+1}\Big\},\end{aligned}\tag{4.6}$$

for $q=O(1)$.

Now, from (4.3) and [5, p. 402], we have $A_0=B_1=(\pi/2e)^{1/2}$. It follows from (4.4) and (4.6) that, for $\lfloor k/2\rfloor=p\gg 1$,

$$\frac{c_0}{c_{2p}}\sim\begin{cases}(-1)^p[\cos(\sqrt{p/2}\,\pi)+(2e/\pi)^{1/2}B_0\sin(\sqrt{p/2}\,\pi)] & \text{if } k=2p,\\ (-1)^p\sqrt{2p}[\sin(\sqrt{p/2}\,\pi)-(2e/\pi)^{1/2}B_0\cos(\sqrt{p/2}\,\pi)] & \text{if } k=2p+1.\end{cases}\tag{4.7}$$

(iii) $2l=k\sin^2\theta$. In this case we assume that

$$0<\theta<\pi/2,\quad \sqrt{k}\,\theta\gg 1,\quad \sqrt{k}(\pi/2-\theta)\gg 1.\tag{4.8}$$

Now $\sin^{2l}t\cos^{k-2l}t$ attains a maximum at $t=\theta$, since

$$\frac{d}{dt}(\sin^{2l}t\cos^{k-2l}t)=\sin^{2l-1}t\cos^{k-2l-1}t[2l\cos^2 t-(k-2l)\sin^2 t],\tag{4.9}$$

and $k-2l=k\cos^2\theta$. Also, from (4.1), we have

$$\begin{aligned}&\int_0^{\pi}\cos(\sqrt{k}\,t)\sin^{2l}t\cos^{k-2l}t\,dt\\ &=\int_{-\theta}^{(\pi/2-\theta)}\{\cos[\sqrt{k}(t+\theta)]+(-1)^k\cos[\sqrt{k}(\pi-t-\theta)]\}\\ &\qquad\cdot\sin^{2l}(t+\theta)\cos^{k-2l}(t+\theta)\,dt.\end{aligned}\tag{4.10}$$

But, for $\tau=O(1)$ and $2l=k\sin^2\theta$, subject to (4.8), it is found that

$$\sin^{2l}(\theta+\tau/\sqrt{k})\cos^{k-2l}(\theta+\tau/\sqrt{k})\sim\sin^{2l}\theta\cos^{k-2l}\theta\, e^{-\tau^2}\quad\text{for } k\gg 1.\tag{4.11}$$

Also [5, p. 402],

$$\int_{-\infty}^{\infty}e^{-\tau^2}\sin\tau\,d\tau=0,\qquad \int_{-\infty}^{\infty}e^{-\tau^2}\cos\tau\,d\tau=\sqrt{\pi}\,e^{-1/4}.\tag{4.12}$$

If we let $\sqrt{k}\,t=\tau$ in (4.10), it follows from (3.16), (4.8), (4.11) and (4.12) that

$$\begin{aligned}c_{2l}&\sim\frac{2\sqrt{\pi}(-1)^{l+1}}{k\sin(\sqrt{k}\,\pi)}e^{-1/4}\left(\frac{2l}{k}\right)^l\left(1-\frac{2l}{k}\right)^{k/2-l}\\ &\quad\cdot\{\cos[\sqrt{k}(\pi-\theta)]+(-1)^k\cos(\sqrt{k}\,\theta)\}.\end{aligned}\tag{4.13}$$

We will now show that this expression for c_{2l} matches with that in (4.4) if $l/\sqrt{k}\ll 1$, corresponding to $\sqrt{k}\,\theta^2\ll 1$. We note that if $\theta=O(k^{-1/3})$, then $\sqrt{k}\,\theta=O(k^{1/6})\gg 1$ and $l/\sqrt{k}=O(k^{-1/6})\ll 1$. Thus,

$$\left(1-\frac{2l}{k}\right)^{k/2-l}\sim e^{-l}\quad\text{if } \frac{l^2}{k}\ll 1.\tag{4.14}$$

Also,

$$\sqrt{2l}=\sqrt{k}\sin\theta\sim\sqrt{k}\,\theta,\quad\text{if } \sqrt{k}\,\theta^3\ll 1.\tag{4.15}$$

Hence, if $\sqrt{l}\gg 1$ and $l/\sqrt{k}\ll 1$, we find from (4.13) that

$$c_{2l}\sim\frac{2\sqrt{\pi}(-1)^{l+1}}{k\sin(\sqrt{k}\,\pi)}\left(\frac{2l}{k}\right)^l e^{-(l+1/4)}[\cos(\sqrt{k}\,\pi-\sqrt{2l})+(-1)^k\cos(\sqrt{2l})].\tag{4.16}$$

The asymptotic behavior of A_{2l} and B_{2l} for $l\gg 1$ may be obtained by substituting $\tau=\sqrt{2l}+x$ in (4.3), with $r=2l$, and using (4.12). It is found that

$$A_{2l}\sim\sqrt{\pi}(2l)^l e^{-(l+1/4)}\cos(\sqrt{2l}),\qquad B_{2l}\sim\sqrt{\pi}(2l)^l e^{-(l+1/4)}\sin(\sqrt{2l}),\tag{4.17}$$

for $l\gg 1$. This shows that (4.4) is consistent with (4.16).

Similarly, it may be shown that the expression (4.13) for c_{2l} matches with that in (4.6) if $l=\lfloor k/2\rfloor-q$ where $\sqrt{q}\gg 1$ and $q/\sqrt{k}\ll 1$. Under these conditions it is found that

$$\begin{aligned}c_{2(\lfloor k/2\rfloor-q)}&\sim\frac{2\sqrt{\pi}(-1)^{\lfloor k/2\rfloor-q+1}}{k\sin(\sqrt{k}\,\pi)}\left(\frac{2q}{k}\right)^q\\ &\quad\cdot e^{-(q+1/4)}\{[1+(-1)^k]\cos(\sqrt{k}\,\pi/2)\cos(\sqrt{2q})\\ &\qquad-[1-(-1)^k]\sqrt{2q/k}\sin(\sqrt{k}\,\pi/2)\sin(\sqrt{2q+1})\}.\end{aligned}\tag{4.18}$$

We have made use of the fact that

$$(2q+1)^{q+1/2}e^{-(q+1/2)}\sim(2q)^{q+1/2}e^{-q}\quad\text{for } q\gg 1.\tag{4.19}$$

In summary, the asymptotic behavior of the coefficients c_{2l}, $l=0,\cdots,\lfloor k/2\rfloor$, in the inversion formula (2.11) is given by (4.4) for $l=O(1)$, by (4.6) for $l=\lfloor k/2\rfloor-q$ and $q=O(1)$, and by (4.13) for $2l=k\sin^2\theta$ where $0<\theta<\pi/2$, and $\sqrt{k}\,\theta\gg 1$ and $\sqrt{k}(\pi/2-\theta)\gg 1$. In addition, $c_{2l+1}=0$. The results indicate that, for $k\gg 1$, the coefficient with the largest magnitude is either c_0 or c_{2p}, where $p=\lfloor k/2\rfloor$. The asymptotic ratio of c_0 to c_{2p} is given by (4.7).

5. Alternate derivation of the inversion formula. We will now give an alternate derivation of the expression (2.12) for the coefficients c_w in the inversion formula (2.11), by considering the transform $g\to\bar{g}$. It follows from (1.3) and (2.10) that

$$(5.1)\qquad \bar{g}(w)=\sum_{r=0}^{k}\alpha_r\sum_{H(z)=w}\;\sum_{H(x,z)=r}f(x).$$

If we use (2.13), and count the number of times that $H(z)=w$, $H(x)=w+r-2s$ and $H(x,z)=r$, we obtain

$$(5.2)\qquad \bar{g}(w)=\sum_{r=0}^{k}\alpha_r\Delta_{rw}(g),$$

where

$$(5.3)\qquad \Delta_{rw}(g)=\sum_{s=0}^{r}\binom{w+r-2s}{r-s}\binom{k-w-r+2s}{s}g(w+r-2s),$$

and $g(l)=0$ if $l<0$ or $l>k$.

Now, from (2.6), the generating function for the Krawtchouk polynomials is

$$(5.4)\qquad \sum_{l=0}^{k}p_l^k(\nu)\xi^l=(1+\xi)^{k-\nu}(1-\xi)^\nu,\qquad \nu=0,\cdots,k.$$

We define

$$(5.5)\qquad g_\nu(l)=\begin{cases}p_l^k(\nu), & l=0,\cdots,k,\\ 0 & \text{otherwise.}\end{cases}$$

Then, from (5.3), if we let $r=s+j$ and $w=l+s-j$, we obtain

$$(5.6)\qquad \sum_{r=0}^{k}\sum_{w=0}^{k}\Delta_{rw}(g_\nu)\eta^r\zeta^w=\sum_{j=0}^{k}\sum_{s=0}^{k-j}\sum_{l=\max(0,j-s)}^{\min(k,k+j-s)}\eta^{s+j}\zeta^{l+s-j}\binom{l}{j}\binom{k-l}{s}p_l^k(\nu).$$

But, if $j>s$ then $\binom{l}{j}=0$ for $0\leqq l<j-s$, and if $s>j$ then $\binom{k-l}{s}=0$ for $k+j-s<l\leqq k$. Hence we may take $0\leqq l\leqq k$ in the sum, and it then follows that

$$\begin{aligned}&\sum_{r=0}^{k}\sum_{w=0}^{k}\Delta_{rw}(g_\nu)\eta^r\zeta^w\\ (5.7)\qquad &=\sum_{l=0}^{k}\sum_{j=0}^{l}\sum_{s=0}^{k-l}\binom{l}{j}\eta^j\zeta^{l-j}\binom{k-l}{s}(\eta\zeta)^s p_l^k(\nu)\\ &=\sum_{l=0}^{k}p_l^k(\nu)(\eta+\zeta)^l(1+\eta\zeta)^{k-l}=[(1+\eta)(1+\zeta)]^{k-\nu}[(1-\eta)(1-\zeta)]^\nu,\end{aligned}$$

from (5.4). Consequently, we have the fundamental result

$$(5.8)\qquad \Delta_{rw}(g_\nu)=p_r^k(\nu)p_w^k(\nu),\qquad r,w=0,\cdots,k.$$

We now let

$$g(l)=\begin{cases}\sum_{\nu=0}^{k} a_\nu p_l^k(\nu), & l=0,\cdots,k,\\ 0 & \text{otherwise},\end{cases} \tag{5.9}$$

where the coefficients a_ν, $\nu=0,\cdots,k$, are to be determined. From (2.7), (5.2), (5.5) and (5.8), since Δ_{rw} is a linear operator, we obtain

$$\bar{g}(w)=\sum_{\nu=0}^{k} a_\nu P(\nu)p_w^k(\nu), \qquad w=0,\cdots,k. \tag{5.10}$$

But, from (5.4), it follows that

$$\sum_{u=0}^{k}\sum_{w=0}^{k} p_u^k(w)p_w^k(\nu)\xi^u=\sum_{w=0}^{k} p_w^k(\nu)(1+\xi)^{k-w}(1-\xi)^w=2^k\xi^\nu. \tag{5.11}$$

This implies the orthogonality relations [4, p. 152, Corollary 18]

$$\sum_{w=0}^{k} p_u^k(w)p_w^k(\nu)=2^k\delta_{u\nu}, \tag{5.12}$$

where δ is the Kronecker symbol. Hence, from (5.10), we find that

$$\sum_{w=0}^{k} p_u^k(w)\bar{g}(w)=2^k a_u P(u). \tag{5.13}$$

Since $p_0^k(\nu)\equiv 1$, from (2.6), it follows from (5.9) and (5.13) that

$$g(0)=\frac{1}{2^k}\sum_{\nu=0}^{k}\frac{1}{P(\nu)}\sum_{w=0}^{k} p_\nu^k(w)\bar{g}(w). \tag{5.14}$$

Since $g(0)=f(0)$, from (2.13), this establishes the expression (2.12) for the coefficients c_w in the inversion formula (2.11).

Acknowledgments. The author is grateful to R. L. Graham for bringing this problem to his attention, to C. L. Mallows for useful discussions, and to P. Diaconis for comments on the first draft of this paper.

REFERENCES

[1] P. Diaconis and R. L. Graham, *The Radon transform on Z_2^k*, Pacific J. Math., to appear.

[2] P. Diaconis, *Projection pursuit for discrete data*, Technical Report No. 148, Dept. Statistics, Stanford Univ., Stanford, CA, 1983.

[3] ———, *The use of group representations in probability and statistics*, Institute of Mathematical Statistics, to appear.

[4] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error Correcting Codes*, North-Holland, Amsterdam, 1977.

[5] W. Magnus, F. Oberhettinger and R. P. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1966.

[6] G. Letac, *Problèmes classiques de probabilité sur un couple de Gelfand*, Analytical Methods in Probability Theory, Lecture Notes in Mathematics 861, D. Dugué, E. Lukacs and V. K. Rohatzi, eds., Springer-Verlag, New York, 1981, pp. 93–120.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

010

ON THE SMALLEST POSITIVE SINGULAR VALUE OF A SINGULAR *M*-MATRIX WITH APPLICATIONS TO ERGODIC MARKOV CHAINS*

JESSE L. BARLOW†

Abstract. Let A be a singular $n \times n$ M-matrix of rank $n-1$ and let B be an $(n-1)\times(n-1)$ nonsingular matrix that results from deleting row j and column k from A. An open question discussed by Harrod and Plemmons [SIAM J. Sci. Stat. Comput., 5 (1984), pp. 453-469] is whether there is a choice of j and k such that $\sigma_{n-1}^{(A)} \approx \sigma_{n-1}^{(B)}$ where $\sigma_{n-1}^{(A)}$ is the smallest positive singular value of A and $\sigma_{n-1}^{(B)}$ is the smallest singular value of B. In this paper, we resolve this conjecture by showing that there is always such a choice.

This conjecture is important when finding the stationary distribution of an ergodic Markov chain. This can be posed as the problem of finding the $n \times 1$ vector p such that $Ap = 0$ and $e^T p = 1$ where $e = (1, \cdots, 1)^T$. Here $A = I - Q^T$ where Q is a row stochastic matrix and p is the stationary distribution vector of the chain. This problem can be reduced to the problem of solving a system of linear equations with coefficient matrix B. If $\sigma_{n-1}^{(A)} \approx \sigma_{n-1}^{(B)}$, then this system of linear equations is about as well conditioned as the original problem.

Key words. M-matrix, singular value decomposition, Markov chains

AMS (MOS) subject classification numbers. 65F, 65G

1. Introduction.

In this paper, we consider the condition of the problem of solving the homogeneous system of linear equations

$$Ap = 0 \tag{1.1a}$$

subject to the constraint

$$e^T p = 1 \tag{1.1b}$$

where $e = (1, \cdots, 1)^T$, A is a singular irreducible $n \times n$ M-matrix with rank $n-1$, and p is an $n \times 1$ vector.

The problem (1.1) arises in the context of finding the stationary distribution of an ergodic Markov chain. This is the problem of finding an $n \times 1$ vector p such that

$$Q^T p = p \tag{1.2}$$

where Q is an $n \times n$ row stochastic matrix. Since Q is an irreducible nonnegative matrix, according to the Perron-Frobenius theory (cf. [27, p. 30], [2, p. 27]), p is the unique positive vector satisfying

$$\sum_{i=1}^{n} p_i = 1. \tag{1.3}$$

This problem has attracted a great deal of interest from a variety of perspectives [18], [12], [22], [20], [25], [26], [16], [17]. The stationary distribution vector p is important in many applications, including queueing networks [17], input-output economic models [2, Chap. 9], [3], and in compartmental analysis tracer models [24], [7]. Such a computation is also related to the discrete Neumann problem in partial differential equations [23].

* Received by the editors May 20, 1985, and in revised form August 6, 1985. This research was supported by the National Science Foundation under contract DCR-8402363 and by the Office of Naval Research under contract N0014-80-0517.

† Department of Computer Science, The Pennsylvania State University, University Park, Pennsylvania 16802.

Let

$$A=I-Q^T \tag{1.4}$$

and $e=(1,\cdots,1)^T$; and the problem (1.2) can be characterized as that of solving (1.1). If A is as in (1.4), then it is a rank $n-1$ singular M-matrix which satisfies

$$e^TA=0. \tag{1.5}$$

The condition of the problem (1.2) has been examined from several points of view concerning the matrix A in (1.4). One common point of view has been through the use of $A^{\#}$, the group inverse of A (cf. [20], [21], [10]). Another has been to consider the eigenvalue of A of second smallest magnitude (cf. [5]). The approach that we will examine will be to use the second smallest singular value of A (cf. [15], [14]).

We consider the matrix A in the form

$$A=P_j\begin{pmatrix} B & y \\ z^T & \alpha_{nn}\end{pmatrix}P_k^T \tag{1.6}$$

where P_j exchanges rows j and n and P_k^T exchanges columns k and n. We assume that A has rank $n-1$ and thus we can guarantee that B is an $(n-1)\times(n-1)$ nonsingular matrix (cf. [15], [10]). The procedure given below as Algorithm 1.1 for solving (1.1) uses the partitioning (1.6).

```
ALGORITHM 1.1.
1. Find the row and column permutation matrices $P_j$ and $P_k^T$.
2. Solve $B\hat{x}=-y$ by $LU$ decomposition or $QR$ decomposition.
3. Let $x=P_k\binom{\hat{x}}{1}$.
4. Set $p=x/(\sum_{i=1}^n x_i)$.
```

The key question that is resolved in this paper is whether there are permutation matrices P_j and P_k^T such that

$$\sigma_{n-1}^{(B)}\approx\sigma_{n-1}^{(A)} \tag{1.7}$$

where $\sigma_{n-1}^{(B)}$ is the smallest singular value of B and $\sigma_{n-1}^{(A)}$ is the smallest positive singular value of A. This conjecture is given in [14] and [10]. If the conjecture is true, then the above procedure yields a system that is about as well conditioned as the original problem (1.2).

We resolve this conjecture in this paper and two related problems: 1) the relationship between $\sigma_{n-1}^{(A)}$ and the condition of the problem (1.1); 2) The choice of the permutation matrices P_j and P_k^T. The problem 1) is resolved in § 2 and the problem 2) is resolved along with the conjecture (1.7) in § 3. A procedure for solving (1.1) that uses these results is given in § 3. This procedure is a modification of Algorithm 1.1.

Throughout this paper we will use the vector norms

$$\|x\|_1=\sum_{i=1}^n |x_i|,\quad \|x\|_2=\sqrt{x^Tx},\quad \|x\|_\infty=\max_{1\leq i\leq n}|x_i|,$$

and the associated induced matrix norms. Unless stated otherwise, the theorems and lemmas in this paper concern singular M-matrices of rank $n-1$ which do not necessarily satisfy the condition (1.5).

2. The condition of the ergodic Markov chain problem. Suppose that we have obtained an approximate solution $\bar{p}$ to (1.1) by some numerical method. Using the techniques of backward error analysis given in [30], we have that $\bar{p}$ satisfies

$$Ap̄ = r, \tag{2.1a}$$

$$e^T\bar{p} = 1 + \varepsilon \tag{2.1b}$$

where r and ε are the residual errors in computation. The perturbation ε on the right side of (2.1b) can always be made approximately equal to the machine unit and does not greatly affect the analysis. We include it for completeness. Although A is singular, (2.1a) is clearly consistent. Thus if we let $\Delta p = \bar{p} - p$ then

$$A\Delta p = 1, \tag{2.2a}$$

$$e^T\Delta p = \varepsilon \tag{2.2b}$$

is consistent. The following theorem gives a bound on Δp in terms of r.

THEOREM 2.1. *Let A be a singular $n \times n$ M-matrix of rank $n-1$. If Δp satisfies* (2.2), *then*

$$\|\Delta p\|_2 \leqq |\varepsilon| + (1+\sqrt{n})\|r\|_2/\sigma_{n-1} \tag{2.3}$$

where σ_{n-1} is the smallest positive singular value of A.

Proof. Let

$$A = U\begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} V^T \tag{2.4}$$

be the singular value decomposition of A where U and V are orthogonal, and $\Sigma = \operatorname{diag}(\sigma_1, \cdots, \sigma_{n-1})$ is an $(n-1)\times(n-1)$ diagonal matrix where $\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_{n-1} > 0$ are the positive singular values of A. Also define

$$\Delta q = V^T\Delta p, \tag{2.5a}$$

$$s = U^T r, \tag{2.5b}$$

$$d = V^T e. \tag{2.5c}$$

The problem (2.2) becomes

$$\begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}\Delta q = s, \tag{2.6a}$$

$$d^T\Delta q = \varepsilon. \tag{2.6b}$$

By consistency of (2.2) we have

$$s = \begin{pmatrix} s_1 \\ 0 \end{pmatrix}$$

where s_1 is an $(n-1)\times 1$ vector. Thus if

$$\Delta p = \begin{pmatrix} \Delta q_1 \\ \rho \end{pmatrix} \tag{2.7}$$

where Δq_1 is an $(n-1)\times 1$ vector and ρ is a scalar, then

$$\|\Delta q_1\|_2 \leqq \|\Sigma^{-1}s_1\|_2 \leqq \|\Sigma^{-1}\|_2\|s_1\|_2 = \|s_1\|_2/\sigma_{n-1} - \|r\|_2/\sigma_{n-1}.$$

The remaining problem is to bound the scalar ρ in (2.7). Let

$$d = \begin{pmatrix} d_1 \\ \delta \end{pmatrix}$$

where d_1 is an $(n-1)\times 1$ vector and δ is a scalar. By definition

$$\delta=(V^Te)_n=v_n^Te$$

where v_n is the singular vector of A such that

$$Av_n=0.$$

By the orthogonality of V we have

$$\|v_n\|_2=1. \tag{2.8}$$

Since the linear space of solutions to (1.1a) has dimension one, $v_n=\kappa p$ where κ is a constant and p is the solution to (1.1). Thus v_n can be chosen to a vector with only positive components which implies that

$$\delta=v_n^Te=\|v_n\|_1. \tag{2.9}$$

Combining (2.8) and (2.9) gives us

$$1\leqq\delta\leqq\sqrt{n}.$$

From the condition (2.6b) we have

$$d^T\Delta q=d_1^T\Delta q_1+\delta\cdot\rho=\varepsilon.$$

Solving for ρ and taking norms we get

$$\begin{aligned}|\rho|&\leqq\frac{|\varepsilon|+\|d_1\|_2\|\Delta q_1\|_2}{|\delta|}\\&\leqq|\varepsilon|+\sqrt{n}\|\Delta q_1\|_2\leqq|\varepsilon|+\sqrt{n}\|r\|_2/\sigma_{n-1}.\end{aligned}$$

Thus by the orthogonality of V

$$\|\Delta p\|_2=\|\Delta q\|_2\leqq|\rho|+\|\Delta q_1\|_2\leqq|\varepsilon|+(1+\sqrt{n})\|r\|_2/\sigma_{n-1}.$$ Q.E.D.

We have shown that the condition of the problem (1.1) depends upon the smallest positive singular value of A. This is equivalent to its dependence on the condition number $\|A\|_2\|A^+\|_2=\sigma_1/\sigma_{n-1}$ where A^+ is the Moore-Penrose inverse of A and σ_1 is its largest singular value.

3. Submatrices of dimension $n-1$. Algorithm 1.1 performs the steps

$$B\hat{x}=-y \tag{3.1a}$$

$$p=P_kx/\|x\|_1 \tag{3.1b}$$

where $x=(\hat{x}^T,1)^T$. It has been shown (cf. [28], [8], [6], [15]) that LU decomposition of an M-matrix without pivoting is a stable process. The step (3.1a), in effect, solves the problem

$$Ax=0, \tag{3.2a}$$

$$x_n=1. \tag{3.2b}$$

Again using classical backward error analysis (cf. [30]) the computed solution to (3.2) using some computational procedure would satisfy

$$A\bar{x}=r, \tag{3.3a}$$

$$\bar{x}_n=1. \tag{3.3b}$$

Thus if $\Delta x = \bar{x} - x$ then it solves

$$A\Delta x = r, \tag{3.4a}$$

$$\Delta x_n = 0. \tag{3.4b}$$

We now consider the solution of (3.4).

LEMMA 3.1. *Let A and r be as in Theorem 2.1, and let*

$$\hat{A} = WAP_k^T$$

where P_k is the permutation matrix that exchanges rows k and n and W is an orthogonal matrix. Let $v_n = (v_{1n}, \cdots, v_{nn})^T$ be the singular vector of A such that

$$Av_n = 0,$$

$$\|v_n\|_2 = 1.$$

If

$$\hat{A}\Delta x = r, \tag{3.5a}$$

$$\Delta x_n = 0 \tag{3.5b}$$

then

$$\|\Delta x\|_2 \leqq (1 + 1/|v_{kn}|)\|r\|_2/\sigma_{n-1} \tag{3.6}$$

and for some choice of k we have

$$\|\Delta x\|_2 \leqq (1+\sqrt{n})\|r\|_2/\sigma_{n-1} \tag{3.7}$$

where σ_{n-1} is the smallest positive singular value of A.

Proof. Let A have the singular value decomposition given in (2.4). The singular value decomposition of $\hat{A}$ is given by

$$\hat{A} = WU\begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} V^T P_k^T.$$

Define

$$s = U^T W^T r,$$

$$\Delta z = V^T P_k^T \Delta x,$$

$$f = V^T P_k^T e_n$$

where $e_n = (0, \cdots, 0, 1)^T$ is the last column of the identity matrix. The problem (3.7) can now be rewritten as

$$\begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} \Delta z = s, \tag{3.8a}$$

$$f^T \Delta z = 0. \tag{3.8b}$$

By consistency

$$s = \begin{pmatrix} s_1 \\ 0 \end{pmatrix}$$

where s_1 is an $(n-1) \times 1$ vector. Therefore if

$$\Delta z = \begin{pmatrix} \Delta z_1 \\ \zeta \end{pmatrix} \tag{3.9}$$

where Δz_1 is an $(n-1)\times 1$ vector and ζ is a scalar, then

$$\|\Delta z_1\|_2 \leqq \|s_1\|_2/\sigma_{n-1} = \|s\|_2/\sigma_{n-1} = \|r\|_2/\sigma_{n-1}. \tag{3.10}$$

The remaining problem is to bound the scalar ζ in (3.9). Let

$$f = \begin{pmatrix} f_1 \\ \gamma \end{pmatrix}.$$

Here

$$\gamma = (V^T P_k^T e_n)_n = (V^T e_k)_n = v_{kn}$$

where $V = (v_{ij})$ is the orthogonal left singular vector matrix in (2.4). Here $v_n = (v_{1n}, \cdots, v_{kn}, \cdots, v_{nn})^T$ is the nth singular vector of A from the hypothesis and thus the nth column of V. The condition (3.8b) gives us that

$$f^T \Delta z = f_1^T \Delta z_1 + \gamma \cdot \zeta = 0.$$

If we solve for ζ and take norms we obtain

$$|\zeta| \leqq \frac{\|f_1\|_2 \|\Delta z_1\|_2}{|\gamma|} \leqq \frac{\|\Delta z_1\|_2}{|v_{kn}|} \leqq \frac{1}{|v_{kn}|} \|r\|_2/\sigma_{n-1}. \tag{3.11}$$

Combining (3.9), (3.10), and (3.11) we obtain

$$\|\Delta x\|_2 = \|\Delta z\|_2 \leqq |\zeta| + \|\Delta z_1\|_2 \leqq (1 + 1/|v_{kn}|)\|r\|/\sigma_{n-1}.$$

If we choose k so that

$$|v_{kn}| = \max_{1 \leqq i \leqq n} |v_{in}| = \|v_n\|_\infty$$

then

$$|v_{kn}| \geqq \frac{1}{\sqrt{n}},$$

which implies

$$\|\Delta x\|_2 \leqq (1 + \sqrt{n})\|r\|_2/\sigma_{n-1}. \qquad \text{Q.E.D.}$$

In (3.6), it is important to avoid permutations P_k^T where v_{kn} is small in magnitude. We can detect this only by solving (1.1a).

We shall need the following result concerning M-matrices.

LEMMA 3.2. *Let A be a singular $n \times n$ matrix. Then there is a positive $n \times 1$ vector c such that*

$$c^T A = 0.$$

For the particular instance of the M-matrices arising out of ergodic Markov chains we have $c = e = (1, \cdots, 1)^T$. The above lemma can be found in [8] and [28].

Theorem 3.1 resolves the open question surrounding (1.7) and is the main theorem of this paper.

THEOREM 3.1. *Let A be a singular M-matrix of rank $n-1$ and let*

$$\hat{A} = P_j A P_k^T$$

where

$$\hat{A} = \begin{pmatrix} B & y \\ z^T & \alpha_{nn} \end{pmatrix}$$

and B is an $(n-1)\times(n-1)$ nonsingular M-matrix. Here P_j is a permutation matrix which exchanges rows j and n and P_k^T is a permutation matrix which exchanges columns k and n. Let $\sigma_{n-1}^{(B)}$ be the smallest singular value of B, and let $\sigma_{n-1}^{(A)}$ be the smallest positive singular value of A. Then, for some choice of j and k, we have

$$\frac{\sigma_{n-1}^{(A)}}{(\sqrt{n}+1)^2} \leqq \sigma_{n-1}^{(B)} \leqq \sigma_{n-1}^{(A)}. \tag{3.12}$$

Proof. The upper bound in (3.12) is proven in [15, p. 459] using the Courant-Fischer minimax theorem (cf. [13, p. 269], [30, pp. 103-104]). Thus we must simply prove the lower bound.

Let $\Delta\hat{x}_1$ be the $(n-1)\times 1$ vector that solves

$$B\Delta\hat{x}_1 = r_1$$

for some $(n-1)\times 1$ vector r_1. By the singular value decomposition theorem, we can always choose r_1 so that

$$\|\Delta\hat{x}_1\|_2 = \|r_1\|_2/\sigma_{n-1}^{(B)}.$$

If we now let

$$\Delta x = \begin{pmatrix} \Delta\hat{x}_1 \\ 0 \end{pmatrix}$$

be an $n\times 1$ vector, then

$$\hat{A}\Delta x = r = \begin{pmatrix} r_1 \\ \rho_n \end{pmatrix}.$$

Let c be the positive vector from Lemma 3.2 such that

$$c^T A = 0.$$

Then we have

$$\hat{c}^T \hat{A}\Delta x = c^T r = 0 \tag{3.13}$$

where

$$\hat{c} = P_j c.$$

Equation (3.13) implies that

$$\hat{c}_n \rho_n = -\sum_{i=1}^{n-1} \hat{c}_i \rho_i, \tag{3.14}$$

where $r = (\rho_1, \rho_2, \cdots, \rho_n)^T$. We can always choose j so that

$$\hat{c}_n = \max_{1\leqq i\leqq n} |\hat{c}_i| \tag{3.15}$$

and therefore

$$|\rho_n| \leqq \sum_{i=1}^{n-1} \frac{\hat{c}_i}{\hat{c}_n} |\rho_i| \leqq \|r_1\|_1. \tag{3.16}$$

A bound on the norm of r is thus given by

$$\|r\|_2 \leqq \|r_1\|_2 + |\rho_n| \leqq \|r_1\|_2 + \|r_1\|_1 \leqq (1+\sqrt{n})\|r_1\|_2. \tag{3.17}$$

Since P_j is an orthogonal matrix, from Lemma 3.1 we can choose k so that

$$\|\Delta x\|_2 = \|\Delta \hat{x}\|_2 \leqq (1+\sqrt{n})\|r\|_2/\sigma_{n-1}^{(A)}.$$

Hence

$$\|\Delta \hat{x}\|_2 = \frac{\|r_1\|_2}{\sigma_{n-1}^{(B)}} \leqq \frac{(1+\sqrt{n})\|r\|_2}{\sigma_{n-1}^{(A)}} \leqq \frac{(1+\sqrt{n})^2\|r_1\|_2}{\sigma_{n-1}^{(A)}}$$

which implies

$$\sigma_{n-1}^{(B)} \geqq \frac{\sigma_{n-1}^{(A)}}{(1+\sqrt{n})^2}. \qquad \text{Q.E.D.}$$

In an informal sense, this theorem states that the problem (1.1) where A is any M-matrix is ill-conditioned if and only if every $(n-1)\times(n-1)$ submatrix of A is also ill-conditioned. This leaves open the method of choosing the row and column pivot indices j and k in Theorem 3.1 in order to solve (1.1) using the strategy in § 1. For the Markov chain problem $c = r = (1, \cdots, 1)^T$; thus, for that case, the choice of j is arbitrary as is stated in the following corollary.

COROLLARY 3.1. *Let A be as in Theorem 3.1 with the condition (1.5). Let $\hat{A}$, P_j, $P_k^T B$, $\sigma_{n-1}^{(A)}$, and $\sigma_{n-1}^{(B)}$ be as in Theorem 3.1. Then for some choice of k and any choice of $j = 1, 2, \cdots, n$ we have (3.12).*

Proof. Simply follow the proof of Theorem 3.1 up until equation (3.13). The vector $c = e = (1, \cdots, 1)^T$ thus for the vector

$$\hat{c} = P_j c$$

the statement (3.15) is true for any choice of $j = 1, 2, \cdots, n$. The remaining statements in the proof of Theorem 3.1 follow. Hence (3.12). Q.E.D.

The last corollary gives a perturbation bound for any arbitrary column permutation for the solution of (3.1).

COROLLARY 3.2. *Let A, $\hat{A}$, P_j, P_k^T, $\sigma_{n-1}^{(A)}$, and $\sigma_{n-1}^{(B)}$ be as in Corollary 3.1. Let $v_n = (v_{1n}, \cdots, v_{nn})^T$ be as in Lemma 3.1. Then for any choice of j and k*

$$\frac{\sigma_{n-1}^{(A)}}{(1/|v_{kn}|+1)(\sqrt{n}+1)} \leqq \sigma_{n-1}^{(B)} \leqq \sigma_{n-1}^{(A)}.$$

Proof. From Lemma 3.1 we have

$$\|\Delta x\|_2 \leqq (1+1/|v_{kn}|)\|r\|_2/\sigma_{n-1}^{(A)}$$

where r is as in the hypothesis of that lemma. Let

$$\Delta x = \begin{pmatrix} \Delta \hat{x}_1 \\ 0 \end{pmatrix}$$

be an $n\times 1$ vector and choose $r = \binom{r_1}{\rho_n}$ so that

$$\|\Delta \hat{x}_1\|_2 = \|r_1\|_2/\sigma_{n-1}^{(B)}.$$

From Corollary 3.1 we have (3.17); thus

$$\|\Delta \hat{x}\|_2 = \frac{\|r_1\|_2}{\sigma_{n-1}^{(B)}} \leqq \frac{(1+1/|v_{kn}|)\|r\|_2}{\sigma_{n-1}^{(A)}} \leqq \frac{(1+1/|v_{kn}|)(1+\sqrt{n})\|r_1\|_2}{\sigma_{n-1}^{(A)}}.$$

We therefore have

$$\frac{\sigma_{n-1}^{(A)}}{(1/|v_{kn}|+1)(\sqrt{n}+1)} \leqq \sigma_{n-1}^{(B)} \leqq \sigma_{n-1}^{(A)}. \qquad \text{Q.E.D.}$$

From [14, p. 43], the vector $v_n = (v_{1n}, \cdots, v_{nn})^T = \beta y$ where $y = (y_1, \cdots, y_n)^T$, y_k is the kth principal minor of A, and β is a constant. In [14, p. 45], an expression for the condition of (1.1a) as an eigenvalue problem is given in terms of the vector y.

Corollary 3.2 yields a stable strategy for solving (1.1) when the condition (1.5) holds as in the ergodic Markov chain problem. We give the strategy using the LU decomposition of B, but a similar strategy can be used with the orthogonal decomposition of B. The strategy is a modification of Algorithm 1.1 and given as Algorithm 3.1.

ALGORITHM 3.1.

1. Compute the factorization

$$LU = B \quad \text{where } A = \begin{pmatrix} B & y \\ z^T & \alpha_{nn} \end{pmatrix}$$

where L is lower triangular and U is upper triangular. From [14], this can be done without pivoting.

2. Let $z = -L^{-1}y$.
3. Solve $U\hat{x}^{(1)} = z$.
4. If $\|\hat{x}^{(1)}\|_\infty < tol$ where $tol \geqq 1$, but tol is not much greater than one, then set

$$x^{(1)} = \begin{pmatrix} \hat{x}^{(1)} \\ 1 \end{pmatrix} \quad \text{and} \quad p = \frac{1}{\|x^{(1)}\|_1} x^{(1)}$$

and quit. Otherwise proceed to step 5.

5. Let $x^{(1)} = (x_1, \cdots, x_n)^T$ from step 4 where $x_n = 1$. Let k be an index such that

$$x_k = \|x\|_\infty = \max_{1 \leqq l \leqq n} x_l.$$

Let $\hat{U}$ be U with the kth column $u^{(k)}$ deleted and let $H = [\hat{U} | z]$. Note that H is upper Hessenberg.

6. Solve $H\hat{x}^{(2)} = u^{(k)}$ using Gaussian elimination with partial pivoting (or equivalently pairwise pivoting). Then

$$x^{(2)} = P_k \begin{pmatrix} \hat{x}^{(2)} \\ 1 \end{pmatrix} \quad \text{and} \quad p = \frac{1}{\|\hat{x}^{(2)}\|_1} x^{(2)}$$

where p is the solution to (1.1).

Pivoting is necessary in step 6 since H is not, in general, an M-matrix, nor is it generalized diagonally dominant. However, the factorization of H by Gaussian elimination with partial pivoting is unconditionally stable [29].

The factorization developed by Algorithm 3.1 is stable. Essentially, we are factoring $\hat{B}$ where

$$A = \begin{pmatrix} \hat{B} & \hat{y} \\ \hat{z}^T & \hat{\alpha}_{nn} \end{pmatrix} P_k^T$$

using the factorization

$$\hat{B} = LH = L\hat{I}\hat{L}\hat{U}$$

where $\hat{I}\hat{L}\hat{U}$ is the LU factorization of H with partial pivoting. Since L comes from

Gaussian elimination on a diagonally dominant matrix, and $\hat{I}\hat{L}\hat{U}$ is the LU factorization of a Hessenberg matrix with partial pivoting, from [29] we have

$$\|L^{-1}\|_1, \|\hat{L}^{-1}\|_1 \leqq n$$

and therefore

$$\|\hat{U}\|_1 \leqq \|L^{-1}\|_1 \|\hat{L}^{-1}\|_1 \|\hat{B}\|_1 \leqq n^2 \|\hat{B}\|_1.$$

Thus there is little enlargement in the elements of $\hat{B}$ which implies stability [1], [4].

Steps 4 and 6 of Algorithm 3.1 make use of Corollary 3.2. Since $x_n = 1$ and $\|v_n\|_2 = 1$, we have $\|x^{(1)}\|_\infty$ is large only if the last component of v_n is small, and thus the lower bound on $\sigma_{n-1}^{(B)}$ is small. Steps 5 and 6 of Algorithm 3.1 are to insure that $\|x^{(2)}\|_\infty = 1$ or is as close as possible to one, thus insuring that $\sigma_{n-1}^{(\hat{B})}$ (the smallest singular value of $\hat{B}$) satisfies the lower bound in equation (3.12).

The same strategy as in Algorithm 3.1 could be used with QR decomposition of $\hat{B}$ except that no pivoting would be necessary in step 6. Also symmetric row and column pivoting can be performed on A before step one to preserve sparsity without disturbing the M-matrix property [2]. Similar procedures to Algorithm 3.1 for deleting and inserting a column into the LU factorization of an M-matrix are discussed in [9].

Acknowledgments. The author would like to thank Professor Robert J. Plemmons with whom he corresponded on this problem. Professor Plemmons also brought the conjecture resolved in Theorem 3.1 to the author's attention and made several suggestions which improved the presentation of this paper.

REFERENCES

[1] J. L. BARLOW, *A Note on monitoring the stability of triangular decomposition of sparse matrices*, Report No. CS-84-02, Dept. Computer Science, The Pennsylvania State Univ., University Park, PA, May 1984; SIAM J. Sci. Stat. Comput., 7 (1986), pp. 166–168.

[2] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[3] F. DUCHIN AND D. B. SZYLD, *Application of sparse matrix techniques to inter-regional input-output analysis*, Economics of Planning, 15 (1979), pp. 142–167.

[4] A. M. ERISMAN AND J. K. REID, *Monitoring the stability of the triangular factorization of a sparse matrix*, Numer. Math., 22 (1974), pp. 183–186.

[5] R. E. FUNDERLIC AND M. T. HEATH, *Linear compartmental analysis of ecosystems*, ORNL/IBP-71/4, Oak Ridge National Laboratory, Oak Ridge, TN, 1971.

[6] R. E. FUNDERLIC AND R. J. PLEMMONS, *LU decomposition of M-matrices by elimination without pivoting*, Linear Algebra and Appl., 41 (1981), pp. 99–110.

[7] R. E. FUNDERLIC AND J. B. MANKIN, *Solution of homogeneous systems of linear equations arising from compartmental models*, SIAM J. Sci. Stat. Comput., 2 (1981), pp. 375–383.

[8] R. E. FUNDERLIC, M. NEUMANN AND R. J. PLEMMONS, *LU Decomposition of generalized diagonally dominant matrices*, Numer. Math., 40 (1982), pp. 57–69.

[9] R. E. FUNDERLIC AND R. J. PLEMMONS, *Updating LU factorizations for computing stationary distributions*, ORNL/Computer Science Dept./TM-208, Computer Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, February 1984; this Journal, 7 (1986), pp. 30–42.

[10] R. E. FUNDERLIC AND C. D. MEYER JR., *Sensitivity of the stationary distribution vector for an ergodic Markov chain*, ORNL-6098, Engineering Physics and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, 1984.

[11] G. H. GOLUB AND C. D. MEYER, JR., *Using the QR factorization and group inversion to compute, differentiate, and estimate the sensitivity of stationary probabilities of Markov chains*, this Journal, pp. 273–281.

[12] G. H. GOLUB AND E. SENATA, *Computation of the stationary distribution of an infinite Markov chain*, Bull. Austral. Math. Soc., 8 (1973), pp. 333–341.

[13] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins Press, Baltimore, 1983.

[14] W. J. HARROD, *Rank modification methods for certain singular systems of linear equations*, Ph.D. Thesis, Univ. Tennessee, Knoxville, TN, 1982.

[15] W. J. HARROD AND R. J. PLEMMONS, *Comparison of some direct methods for computing stationary distributions of Markov chains*, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 453-469.

[16] J. J. HUNTER, *Generalized inverses and their application to applied probability problems*, Linear Algebra Appl., 45 (1982), pp. 157-198.

[17] L. C. KAUFMAN, *Matrix methods for queueing problems*, SIAM J. Sci. Stat. Comput., 4 (1983), pp. 525-552.

[18] J. G. KEMENY AND J. L. SNELL, *Finite Markov Chains*, Van Nostrand, Princeton, NJ, 1960.

[19] J. G. KEMENY, *Generalization of a fundamental matrix*, Linear Algebra Appl., 38 (1981), pp. 193-226.

[20] C. D. MEYER JR., *The role of the group inverse in the theory of finite Markov chains*, SIAM Rev., 17 (1975), pp. 443-464.

[21] ———, *The condition of a finite Markov chain and perturbation bounds for the limiting probabilities*, this Journal, 1 (1980), pp. 273-283.

[22] C. C. PAIGE, P. H. STYAN AND P. G. WACHTER, *Computation of the stationary distribution of a Markov chain*, J. Statist. Comput. Simul., 4 (1975), pp. 173-186.

[23] R. J. PLEMMONS, *Regular splittings and the discrete Neumann problem*, Num. Math., 25 (1976), pp. 153-161.

[24] C. W. SHEPPARD AND A. S. HOUSEHOLDER, *The mathematical basis of the interpretation of tracer experiments in closed steady-state systems*, J. Appl. Phys., 22 (1951), pp. 510-520.

[25] W. J. STEWART, *A comparison of numerical techniques in Markov modelling*, Comm. Assoc. Comput. Mach., 21 (1978), pp. 144-151.

[26] G. W. STEWART, *Computable error bounds for aggregated Markov chains*, J. Assoc. Comput. Mach., 30 (1983), pp. 271-285.

[27] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[28] R. S. VARGA AND D. CAI, *On the LU factorization of M-matrices*, Num. Math., 38 (1981), pp. 179-192.

[29] J. H. WILKINSON, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8 (1961), pp. 281-330.

[30] ———, *The Algebraic Eigenvalue Problem*, Oxford Univ. Press, London, 1965.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

011

CURVES ON S^{n-1} THAT LEAD TO EIGENVALUES OR THEIR MEANS OF A MATRIX*

MOODY T. CHU†

Abstract. This paper discusses two dynamical systems on the unit sphere S^{n-1} in $\mathbb{R}^n$ space, each defined in terms of a real square matrix M. The solutions of these systems are found to converge to points which provide essential information about eigenvalues of the matrix M. It is shown, in particular, how the dynamics of the second flow is analogous to that of the Rayleigh quotient iterations.

Key words. dynamical systems, Lyapunov's theorem, Rayleigh quotient iterations, stability

1. Introduction. Consider the following autonomous differential system:

$$\frac{dy}{dt} = F(y) \tag{1.1}$$

where $y(t) \in \mathbb{R}^n$, F is continuous wherever it is defined and satisfies

$$F(y) = \|y\| \cdot F(y/\|y\|) \qquad (\|\cdot\| \text{ means the 2-norm}) \tag{1.2}$$

for every $y \neq 0$. For any nonzero solution $y(t)$ of (1.1) if we define

$$u(t) = y(t)/\|y(t)\|, \tag{1.3}$$

then $u(t)$ is a smooth flow on S^{n-1} and satisfies the system

$$\frac{du}{dt} = F(u) - \langle u, F(u)\rangle \cdot u. \tag{1.4}$$

Since the unit sphere S^{n-1} is a compact set in the space $\mathbb{R}^n$, we expect the global dynamics of the vector field (1.4) on this unit sphere to have some kinds of convergence properties. We are particularly interested in the case when

$$F(y) = \mathcal{M}(y) \cdot y \tag{1.5}$$

where $\mathcal{M}(y)$ is a square matrix satisfying

$$\mathcal{M}(y) = \mathcal{M}(y/\|y\|). \tag{1.6}$$

More specifically, given a constant matrix M in $\mathbb{R}^{n\times n}$, we shall use M to construct a mapping $\mathcal{M}$ satisfying (1.6) and to study the resulting dynamics. It turns out that all our constructions of $\mathcal{M}$ generate curves leading to points which give information about eigenvalues of M.

Our presentation is organized as follows. In § 2 we discuss the simplest possible choice of (1.6) and show that the dynamics of the resulting system is closely related to that of the classical power method. In § 3, the major part of this paper, we first construct a nontrivial mapping $\mathcal{M}(y)$. It is shown that the flow of the corresponding system of (1.4) satisfies a nice monotone residue property. Using this property, we are able to analyze the global dynamics of the system which turns out to be very analogous to that of the Rayleigh quotient iteration method. The paper is concluded with some remarks in § 4.

* Received by the editors May 24, 1985, and in revised form August 6, 1985.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205.

2. Continuous power method. In this section we first consider the mapping $\mathcal{M}(y) \equiv M$. The system

$$\frac{du}{dt} = M \cdot u - \langle u, M \cdot u \rangle \cdot u \tag{2.1}$$

has equilibrium points only at eigenvectors of M. It is also clear from (1.3) that the exact solution of (2.1) is given by

$$u(t) = e^{M \cdot t} \cdot u_0 / \|e^{M \cdot t} \cdot u_0\|. \tag{2.2}$$

To analyze the global behavior of the flow (2.2), we decompose M into its Jordan canonical form, i.e., $M = S \cdot J \cdot S^{-1}$, where

$$J = \bigoplus_{k=1}^{m} J_{n_k}(\lambda_k)$$

and $J_{n_k}(\lambda_k)$ is the $n_k \times n_k$ Jordan block associated with the eigenvalue λ_k.

DEFINITION 2.1. We shall say that λ_i *dominates* λ_j if and only if $\operatorname{Re} \lambda_i \geqq \operatorname{Re} \lambda_j$.

Partition the matrix S as

$$S = [S^{(1)}, \cdots, S^{(m)}]$$

where each $S^{(k)}$ is an $n \times n_k$ submatrix of S, and suppose $S^{(k)} = [s_{ij}^{(k)}]$. Then (2.2) can be rewritten as

$$u(t) = S \cdot e^{J \cdot t} \cdot S^{-1} \cdot u_0 / \|S \cdot e^{J \cdot t} \cdot S^{-1} \cdot u_0\|. \tag{2.3}$$

Note that the components of $\hat{u}_0 = S^{-1} \cdot u_0$ are the coordinates of the initial vector u_0 in terms of column vectors of S. Let $\hat{u}_0$ be partitioned as

$$\hat{u}_0^T = [\hat{u}_0^{(1)}, \cdots, \hat{u}_0^{(m)}],$$

where each $\hat{u}_0^{(k)}$ is a $1 \times n_k$ row vector. We shall assume the "generic" condition on the initial vector u_0, i.e.

$$\hat{u}_0^{(1)} \neq 0. \tag{2.4}$$

Since

$$e^{J_{n_k}(\lambda_k)t} = e^{\lambda_k t} \begin{bmatrix} 1 & t & & \cdots & \dfrac{t^{n_k - 1}}{(n_k - 1)!} \\ & 1 & t & \cdots & \vdots \\ & & & \ddots & \\ & & & & t \\ & 0 & & & 1 \end{bmatrix} \tag{2.5}$$

the ith component $u_i(t)$ of the solution $u(t)$ is given by

$$u_i(t) = \sum_{k=1}^{m} e^{\lambda_k t} \sum_{j=1}^{n_k} s_{ij}^{(k)} P_j^{(k)} t \Bigg/ \left\{ \sum_{i=1}^{n} \left| \sum_{k=1}^{m} e^{\lambda_k t} \sum_{j=1}^{n_k} s_{ij}^{(k)} P_j^{(k)}(t) \right|^2 \right\}^{1/2} \tag{2.6}$$

where $P_j^{(k)}t$ is the jth component of the vector

$$e^{-\lambda_k t}\cdot e^{J_{n_k}(\lambda_k)}\cdot \hat{u}_0^{(k)T},$$

and thus is a polynomial in t with degree not higher than n_k-j. From (2.6), we are able to make the following observations [4].

THEOREM 2.1. (1) *If the matrix M has a real eigenvalue, say λ, which dominates any other eigenvalues, then under the assumption* (2.4), *the solution flow $u(t)$ of* (2.1) *converges to the corresponding eigenvector of λ.*

(2) *If λ is complex-valued, then the solution flow $u(t)$ is attracted to and oscillating around the 2-dimensional circle formed by the intersection of S^{n-1} and the 2-dimensional subspace spanned by the real and the imaginary parts of the corresponding eigenvector of λ.*

Remark. It is worth noting that the convergence property in (1) is independent of the order of the corresponding elementary divisor.

Suppose now that M is nonsingular and we choose $\mathcal{M}(y)\equiv \ln M$, the logarithm of M. Then (2.2) becomes

$$u(t)=M^t\cdot u_0/\|M^t\cdot u_0\|, \tag{2.7}$$

where by M^t we mean the matrix exponential $e^{t\cdot(\ln M)}$. It is clear that the evaluations of (2.7) at integer times correspond to the classical power method [7] applied to the matrix M. It is for this reason that the system (2.1) is called "the continuous power method."

3. Continuous analogue of Rayleigh quotient iterations. In this section we define $\mathcal{M}$ to be the mapping

$$\mathcal{M}(y)=[(M-\rho(y))^T\cdot(M-\rho(y))]^{-1}, \tag{3.1}$$

where $\rho(y)$ is the Rayleigh quotient of y with respect to M, i.e.,

$$\rho(y)=\langle y, My\rangle/\langle y, y\rangle, \tag{3.2}$$

$M-\rho(y)$ represents the matrix algebra $M-\rho(y)\cdot I$ and $(M-\rho(y))^T$ is the transpose of $M-\rho(y)$. Let $\sigma(M)$ denote the spectrum of M. Apparently the system

$$\frac{du}{dt}=\mathcal{M}(u)\cdot u-\langle u, \mathcal{M}(u)\cdot u\rangle\cdot u \tag{3.3}$$

becomes singular at any point of Γ if

$$\Gamma=\{u\in S^{n-1};\ \rho(u)\in\sigma(M)\}.$$

Henceforth, we shall consider (3.3) only in the domain $D=S^{n-1}-\Gamma$ which is open with respect to the induced topology on S^{n-1}. Let us define the residue function

$$r(t)=\|(M-\rho(u(t))\cdot u(t)\| \tag{3.4}$$

along any solution flow $u(t)$ of (3.3). The following interesting monotone residue property is found to be important in later analysis.

THEOREM 3.1. *Along any solution $u(t)$ of* (3.3), *the residue function $r(t)$ satisfies $dr/dt\leqq 0$, i.e., $r(t)$ is nonincreasing. The equality holds if and only if $u(t)$ is an eigenvector of $\mathcal{M}$, i.e., $u(t)$ is an equilibrium point of* (3.3).

Proof. From (3.4), we know

$$
\begin{aligned}
2r\frac{dr}{dt} &= \frac{d}{dt}\langle u, \mathcal{M}^{-1}u\rangle \\
&= \left\langle \frac{du}{dt}, \mathcal{M}^{-1}u\right\rangle + \left\langle u, -\frac{d\sigma}{dt}[(M-\rho)+(M-\rho)^T]u + \mathcal{M}^{-1}\frac{du}{dt}\right\rangle \\
&= 2\left\langle \frac{du}{dt}, \mathcal{M}^{-1}u\right\rangle \qquad (\text{since } \langle u, (M-\rho)u\rangle = 0 \text{ and } \mathcal{M} \text{ is symmetric}) \\
&= 2\langle \mathcal{M}u - \langle u, \mathcal{M}u\rangle u, \mathcal{M}^{-1}u\rangle \\
&= 2\{1 - \langle u, \mathcal{M}u\rangle\langle u, \mathcal{M}^{-1}u\rangle\}.
\end{aligned}
$$

Since $\mathcal{M}$ is always positive definite so long as $u(t)$ exists, by the well-known Kantorovich inequality [6],

$$1 \leqq \langle u, \mathcal{M}u\rangle\langle u, \mathcal{M}^{-1}u\rangle, \tag{3.5}$$

we obtain the result $dr/dt \leqq 0$. It is also known from [6], that the equality of (3.5) holds if and only if u is an eigenvector of $\mathcal{M}(u)$. Hence the theorem is proved.

Let E be the collection of equilibrium points of (3.3), i.e.

$$E = \{u^* \in D;\ u^* \text{ is an eigenvector of } \mathcal{M}(u^*)\}.$$

With the residue function $r(t)$ on hand, we may apply Lyapunov's theorem to investigate the global behavior of $u(t)$ in D.

THEOREM 3.2. *Given an arbitrary matrix M in $\mathbb{R}^{n\times n}$, let $u(t)$ be the solution of the corresponding system* (3.3) *with initial value $u_0 \in D$. Then $u(t) \in S^{n-1}$. Furthermore, for $t \geqq 0$, the following three mutually exclusive cases are the only possible manners in which the trajectory of $u(t)$ can behave.*

(1) *$u(t)$ approaches the singular set Γ in finite time.*

(2) *$u(t)$ converges to an eigenvector of M as t goes to infinity.*

(3) *$u(t)$ has its ω-limit set contained in E.*

Proof. It is clear from (1.4) and (3.3) that $u(t) \in S^{n-1}$. If the closure of the positive semiorbit of $u(t)$ is properly contained in D, then $u(t)$ has nonempty ω-limit set Ω in D. By Theorem 3.1, Ω can contain only points of E.

It is obvious that if either of the first two cases of the above theorem happens, then the corresponding $\rho(t)$ approaches an eigenvalue of M. The following two theorems explain, respectively, how the last case should be interpreted for real symmetric and normal matrices.

THEOREM 3.3. *Suppose that the matrix $M \in \mathbb{R}^{n\times n}$ is symmetric and that its spectrum is simple. Then*

(1) *The set E consists of all bisections of pairs of eigenvectors of M. In each 2-dimensional subspace spanned by a pair of eigenvectors, there are only four such points.*

(2) *If the last case of Theorem* 3.2 *happens, then $\rho(t)$ converges to the mean of the pair of eigenvalues whose corresponding eigenvectors are bisected by that equilibrium point.*

(3) *Points in E are unstable. In fact, they are saddles.*

Proof. It is not difficult to see that the vector field (3.3) is invariant under orthogonal transformation. So, without loss of generality, we may assume that M is a diagonal matrix. Obvious $\mathcal{M}$ can not be defined at any of the standard unit vectors.

If at a point $u^* \in E$ the diagonal matrix $\mathcal{M}(u^*)$ has an eigenvector other than any of the standard unit vectors, then it must be that $\mathcal{M}(u^*)$ has equal diagonal elements (eigenvalues) so that an entire 2-dimensional coordinate plane becomes an eigenspace.

This is possible only if $M-\rho(u^*)$ has equal module but opposite sign eigenvalues. In other words, $\rho(u^*)$ must be the midpoint of a pair of diagonal elements of M and u^* must be the bisector of the corresponding coordinate axes. There are exactly four such points in each coordinate plane.

To examine the stability of these equilibrium points, we linearize system (3.3) at an equilibrium point, say $u_{st}^*=(0,\cdots,0,1/\sqrt{2},0,\cdots,0,1/\sqrt{2},0,\cdots,0)$. Suppose $M=\operatorname{diag}(\sigma_1,\cdots\sigma_n)$. Then a tedious but straightforward calculation shows that the coefficient matrix $\partial\dot{u}_i/\partial u_j$ of the variation equation is given by

$$\frac{\partial \dot{u}_i}{\partial u_j}=\begin{cases}\dfrac{20\sigma_t-4\sigma_s}{(\sigma_s-\sigma_t)^2} & \text{if } (i,j)=(s,t),\\[2ex] \dfrac{12\sigma_s+4\sigma_t}{(\sigma_s-\sigma_t)^2} & \text{if } (i,j)=(s,s),\\[2ex] \dfrac{12\sigma_t+4\sigma_s}{(\sigma_t-\sigma_s)^3} & \text{if } (i,j)=(t,t),\\[2ex] \dfrac{20\sigma_s-4\sigma_t}{(\sigma_t-\sigma_s)^3} & \text{if } (i,j)=(t,s),\\[2ex] 0 & \text{otherwise.}\end{cases} \tag{3.6}$$

This matrix has eigenvalues $16/(\sigma_s-\sigma_t)^2$, $-8/(\sigma_s-\sigma_t)^2$ and $4\{1/(2\sigma_i-\sigma_s-\sigma_t)^2-1/(\sigma_s-\sigma_t)^2\}$ for $i\neq s$ or t. Therefore, in the invariant (s,t)-coordinate plane the equilibrium point u_{st}^* is a saddle point. Furthermore, since the unstable eigenvector $[0,\cdots,0,1/\sqrt{2},0,\cdots,0,-1/\sqrt{2},0,\cdots,0]^T$ is always tangential to the sphere S^{n-1}, u_{st}^* is an unstable point on S^{n-1}.

THEOREM 3.4. *Suppose that the matrix $M\in\mathbb{R}^{n\times n}$ is normal and that its spectrum is $\lambda_1\pm i\lambda_2,\lambda_3,\cdots,\lambda_n$ where all λ_i's are real, $\lambda_1,\lambda_3,\cdots,\lambda_n$ are distinct and $\lambda_2\neq 0$. Then*

(1) *Points in E can be classified into three categories denoted as set A, B, and C respectively, where*

$A=\{$*point which bisects a pair of real eigenvectors of* $M\}$;

$B=\{u^*\in D$; *there exists at least one real eigenvalue of M, say* λ_i, *such that* $|\lambda_1+i\lambda_2-\rho(u^*)|=|\lambda_i-\rho(u^*)$, *and* $\rho(u^*)$ *lies between* λ_1 *and* $\lambda_i\}$;

$C=\{$*point which is in the intersection of* S^{n-1} *and the 2-dimensional subspace spanned by the real and the imaginary parts of the eigenvector associated with the eigenvalue* $\lambda_1+i\lambda_2\}$.

(2) *The stability properties of points in E are described as follows.*

(a) *Points in A are isolated and are always unstable.*

(b) *The set B may be empty. If it is not empty, then points in B are not isolated and are always unstable.*

(c) *The circular set C may or may not be stable depending on the position of* $\lambda_1+i\lambda_2$ *relative to other real eigenvalues.*

Proof. Again, without loss of generality, we may assume that M is of the form

$$M=\begin{bmatrix}\lambda_{1'} & \lambda_{2'} & & & 0\\ -\lambda_{2'} & \lambda_{1'} & & & \\ & & \lambda_{3'} & & \\ & & & \ddots & \\ 0 & & & & \lambda_n\end{bmatrix}.$$

Therefore, for any $u \in D$ the matrix $\mathcal{M}(u)$ is always a diagonal matrix with diagonal elements

$$1/[\lambda_1-\rho(u))^2+\lambda_2^2], \quad 1/[\lambda_1-\rho(u))^2+\lambda_2^2], \quad 1/((\lambda_3-\rho(u))^2, \quad \cdots, \quad 1/(\lambda_n-\rho(u))^2.$$

Apparently, points in C are in D and are eigenvectors of $\mathcal{M}$ and, hence, are in E. The fact that points in A are also in E follows from the same arguments as in Theorem 3.3(1). If the system (3.3) has any equilibrium point u^* other than those in A or C, then it must be such that $(\lambda_1-\rho(u^*))^2+\lambda_2^2=(\lambda_i-\rho(u^*))^2$ for some (but at least one) $i \geqq 3$. In other words, $\rho(u^*) \in \mathbb{R}$ must be equidistance from $\lambda_1 \pm i\lambda_2$ and λ_i. This is a necessary condition for $u^* \in B$.

The isolation and stability properties of points in A follow from the same arguments as in Theorem 3.3(3). To see (2b), we first realize that the matrix M is invariant under $(1,2)$-plane rotations. Therefore, dynamics of (3.3) in the $(1, i)$-plane determine the entire dynamics in the $(1,2,i)$-space just by plane rotations. Since in the $(1,i)$-plane we have

$$\begin{aligned} \frac{du_1}{dt}&=\left[\frac{1}{(\lambda_1-\rho(u))^2-\lambda_2^2}-\frac{1}{(\lambda_i-\rho(u))^2}\right]u_i^2u_1, \\ \frac{du_i}{dt}&=\left[\frac{1}{(\lambda_i-\rho(u))^2}-\frac{1}{(\lambda_1-\rho(u))^2+\lambda_2^2}\right]u_1^2u_i \end{aligned} \tag{3.7}$$

with

$$\rho(u)=\lambda_1u_1^2+\lambda_iu_i^2=(\lambda_1-\lambda_i)u_1^2+\lambda_i, \tag{3.8}$$

it is clear that $u^* \in S^{n-1} \cap (1,i)$-plane is an equilibrium point if and only if

$$(\lambda_1-\rho(u^*))^2+\lambda_2^2=(\lambda_i-\rho(u^*))^2 \quad \text{and} \quad 0<\frac{\rho(u^*)-\lambda_i}{\lambda_1-\lambda_i}\leqq 1. \tag{3.9}$$

It is plain to see from (3.8) that (3.9) is possible if and only if

$$|\lambda_1-\lambda_i| \geqq |\lambda_2|. \tag{3.10}$$

In this case, for any point $u \in D$ near u^*, $du_1/dt \cdot u_1>0$ if and only if $|u_1|>|u_1^*|$. Thus we have proved (2b) that every point in B is unstable. We have also shown that if (3.10) holds for all $i \geqq 3$, then the circular set C is a stable center manifold. Finally, if $|\lambda_1-\lambda_i|<|\lambda_2|$ then $du_1/dt \cdot u_1<0$ and $du_i/dt \cdot u_i>0$. In this case, the circular set C is unstable.

Remark. We now explain the analogy between the system (3.3) and the classical Rayleigh quotient iterations. First, recall that the Rayleigh quotient iterations generates a sequence of unit vectors $\{x_k\}$ from a given unit vector x_0 as follows:

$$\begin{aligned} &\mu_k=\rho(x_k), \\ &z_{k+1}=(M-\mu_k)^{-1}\cdot x_k, \qquad x_{k+1}=z_{k+1}/\|z_{k+1}\|, \qquad k=0,1,\cdots. \end{aligned} \tag{3.11}$$

If the system (3.3) is defined piecewise in the way that for all $t \in [k, k+1]$,

$$\mathcal{M}(u(t)) \equiv \ln[(M-\rho_k)]^{-1} \tag{3.12}$$

where $\rho_k=\rho(u(k))$ and $u(t)$ is the solution of the corresponding system (3.3), it can be shown that

$$u(k+1)=(M-\rho_k)^{-1}\cdot u(k)/\|(M-\rho_k)^{-1}\cdot u(k)\|. \tag{3.13}$$

In other words, the solution $u(t)$ of (3.3), with $\mathcal{M}$ defined by (3.12), is a "continuous" generalization of the sequence $\{x_k\}$ of (3.8), provided $x_0 = u(0)$. In practice, however, it is unadvisable to compute the matrix logarithm. So we want to replace (3.12) by a "similar" but much "easier-to-handle" alternative. Toward the end, we recall that the major theme in Rayleigh quotient iterations is to speed up the convergence, and that this theme is carried out by making $M - \mu_k$ "nearly singular" so that $(M - \mu_k)^{-1}$ has a most dominant eigenvalue (see Definition 2.1). Consider now that generally we may suppose that M is diagonalizable, say

$$M = S \cdot D \cdot S^{-1}, \tag{3.14}$$

where $D = \operatorname{diag}\{d_1, \cdots, d_n\}$. Thus

$$\begin{aligned} \ln[\mu - \rho_k)^{-1}] &= S \cdot \ln[(D - \rho_k)^{-1}] \cdot S^{-1} \\ &= S \cdot \operatorname{diag}\{\ln[(d_1 - \rho_k)^{-1}], \cdots, \ln[(d_n - \rho_k)^{-1}]\} \cdot S^{-1}. \end{aligned} \tag{3.15}$$

Observe that the "qualitative behavior" of the two scalar functions $\ln[(z-c]^{-1}]$ and $(z-c)^{-2}$ are very "similar" near the singular point c. (Quantitatively, however, the latter blows up faster than the former.) So our first thought is to replace $\ln[(M - \rho_k)^{-1}]$ by $(M - \rho_k)^{-2}$. After deliberations upon other concerns, such as the applicability of the Kantorovich inequality (3.5) and the continuity of (3.2), we finally decide to replace $\ln[(M - \rho_k)]^{-1}$ by $[(M - \rho(u(t)))^T \cdot (M - \rho(u(t)))]^{-1}$. The feasibility of this substitution is evidenced by the fact that most of the theorems concerning the dynamics of the system (3.3) are almost parallel with those of the classical Rayleigh quotient iterations. Interested readers can refer to [7] and [8] to compare these properties.

4. More remarks.

(1) Suppose, instead of (3.1), one chooses

$$\mathcal{M}(y) = (M - \rho(y))^{-1}. \tag{4.1}$$

Then Theorem 3.1 no longer holds. However, if M is symmetric, then we have the surprising but nice feature that $\rho(t)$ is monotone increasing along any solution flow $u(t)$ on S^{n-1}. Indeed,

$$\frac{d\rho}{dt} \equiv 2. \tag{4.2}$$

This can be seen from

$$\begin{aligned} \frac{d\rho}{dt} = 2\left\langle \frac{du}{dt}, Mu \right\rangle &= 2\langle \mathcal{M}u - \langle u, \mathcal{M}u\rangle u, Mu\rangle \\ &= 2\langle \mathcal{M}u, Mu - \langle u, Mu\rangle u\rangle = 2\langle \mathcal{M}u, \mathcal{M}^{-1}u\rangle = 2. \end{aligned}$$

What this suggests for symmetric eigenvalue problems is that if one starts randomly on S^{n-1} and follows carefully the curve determined by (3.3) with $\mathcal{M}$ defined by (4.1), then one is guaranteed that the corresponding $\rho(t)$ approaches the eigenvalue which is immediately next to the right of the Rayleigh quotient of the starting value. By a suitable deflation technique or by restarting, we may find all eigenvalues.

(2) Note that both Theorem 3.1 and Theorem 3.2 are true for general real matrices. Hence the global convergence of the flow $u(t)$ is always expected. The troublesome set E always has measure zero in $\mathbb{R}^n$ except for trivial M. Moreover, unless under very special circumstances, (e.g. $|\lambda_1 - \lambda_i| \geqq |\lambda_2|$ for all $i \geqq 3$) we have seen that points in E are often unstable (a proof for a general matrix is needed!). All of these observations

seem to suggest that if one starts randomly on S^{n-1}, then with probability one the corresponding $\rho(t)$ of the solution of the system (3.3) approaches an eigenvalue.

(3) In [1] and [3], we have pointed out that the system (2.1) alone determines the entire dynamics of the so called Toda lattice

$$\frac{dx}{dt}=[X,\Pi_0X]=X\cdot\Pi_0X-\Pi_0X\cdot X \tag{4.3}$$

where X is an $n\times n$ square matrix and $\Pi_0X=X^- -(X^-)^T$ with X^- the strictly lower triangular part of X. It has also been pointed out that the Toda flow is closely related to the unshifted QR-algorithm. See, for example, [2], [5], and [9]. Unfortunately, it is not at all clear how (4.3) should be modified to model the shifted QR-algorithm. While it has been observed but no proof has yet been given that the shifted QR-algorithm converges almost always for nonnormal matrices, a proof of (almost) global convergence of Rayleigh quotient iterations is required. We hope that the similarity between properties of the system discussed in this paper and those of Rayleigh quotient iterations will stimulate further efforts in that direction.

Acknowledgment. The author thanks his colleague, Professor Larry Norris, for making many helpful comments.

REFERENCES

[1] M. T. CHU, *On the global convergence of the Toda lattice for real normal matrices and its application to the eigenvalue problem*, SIAM J. Math. Anal., 15 (1984), pp. 98-104.

[2] ———, *The generalized Toda lattice, the QR-algorithm and the center manifold theory*, this Journal, 5 (1984), pp. 187-201.

[3] ———, *Asymptotic analysis of the Toda lattice on diagonalizable matrices*, Nonlinear Anal., TMA, 9 (1985), pp. 193-202.

[4] ———, *Continuous power method*, manuscript.

[5] P. DEIFT, T. NANDA AND C. TOMEI, *Differential equations for the symmetric eigenvalue problems*, SIAM J. Numer. Anal., 20 (1983), pp. 1-22.

[6] M. MARCUS AND H. MINC, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, 1964.

[7] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[8] ———, *The Rayleigh quotient iteration and some generalizations for nonnormal matrices*, Math. Comp., 28 (1974), pp. 679-693.

[9] W. W. SYMES, *The QR-algorithm and scattering for the finite nonperiodic Toda lattice*, Physics, 4D (1982), pp. 275-280.

CONVEXITY IN GRAPHS AND HYPERGRAPHS*

MARTIN FARBER† AND ROBERT E. JAMISON‡

Abstract. We study several notions of abstract convexity in graphs and hypergraphs. In each case, we obtain analogues of several classical results, including the Minkowski-Krein-Milman theorem, Caratheodory's theorem and Tietze's convexity theorem. In addition, our results yield new characterizations of the classes of chordal gaphs, strongly chordal graphs, Ptolemaic graphs and totally balanced hypergraphs.

Key words. convex geometry, geodesic convexity, chordal graphs, totally balanced hypergraphs

AMS(MOS) subject classifications. 056, 52

1. Introduction. An *alignment* on a finite set X is a family $\mathscr{L}$ of subsets of X (to be considered convex sets), which is closed under intersection and which contains both X and the empty set. The pair $(X, \mathscr{L})$ is called an aligned space. The smallest member of $\mathscr{L}$ containing a set $S \subseteq X$ is the *hull* of S, denoted $\mathscr{L}(S)$. An element p of a set $Y \in \mathscr{L}$ is an *extreme point* of Y if $Y \backslash \{p\} \in \mathscr{L}$. A *convex geometry* (antimatroid [19]) on a finite set is an aligned space satisfying the following additional property:

MINKOWSKI-KREIN-MILMAN PROPERTY. *Every convex set is the hull of its extreme points.*

Equivalently, a convex geometry is an aligned space satisfying:

ANTIEXCHANGE PROPERTY. *For any convex set K and two distinct points $x, y \notin K$, $x \in \mathscr{L}(K \cup \{y\})$ implies $y \notin \mathscr{L}(K \cup \{x\})$.*

The following fundamental result follows immediately from [9, Lemma 3.2], or [19, Thm. 1].

THEOREM 1.1. *If $(X, \mathscr{L})$ is a convex geometry, then $K \in \mathscr{L}$ if and only if there is an ordering $p_1, p_2, \cdots, p_m$ of $X \backslash K$ such that p_i is an extreme point of $K \cup \{p_i, p_{i+1}, \cdots, p_m\}$, for $i = 1, 2, \cdots, m$.*

Numerous classes of graphs (e.g., forests, chordal graphs and strongly chordal graphs) can be characterized in the following general way:

> G is a member of $\mathscr{X}$ if there is an ordering $v_1, v_2, \cdots, v_n$ of $V(G)$ such that v_i satisfies property $*$ in the subgraph induced by $\{v_i, v_{i+1}, \cdots, v_n\}$.

For example, if property $*$ is "has degree 0 or 1", then $\mathscr{X}$ is the class of forests, and if property $*$ is "is simplicial" then $\mathscr{X}$ is the class of chordal graphs [7], [27]. Theorem 1.1. suggests that such classes of graphs might be related to convex geometries. Indeed, if property $*$ is maintained by vertices under taking induced subgraphs, and $\mathscr{L}(G)$ is the collection of subsets of $V(G)$ obtainable from $V(G)$ by sequentially deleting vertices satisfying property $*$, then $(V(G), \mathscr{L}(G))$ is a convex geometry if and only if $G \in \mathscr{X}$. In such cases, it is natural to ask for internal descriptions of convex sets and hulls of arbitrary sets. For example, if property $*$ is "has degree 0 or 1" and G is a tree, then a set $S \subseteq V(G)$ is convex if and only if it induces a connected subgraph.

* Received by the editors April 5, 1984, and in revised form August 9, 1985.

† Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. Present address, AT&T Bell Laboratories, Holmdel, New Jersey 07733. The research of this author was supported in part by the Natural Sciences and Engineering Research Council of Canada.

‡ Department of Mathematical Sciences, Clemson University, Clemson, South Carolina 29631. The research of this author was supported in part by the National Science Foundation under grant ISP-8011451 (EPSCoR).

On the other hand, given a collection $\mathscr{L}$ of subsets of $V(G)$, one can ask when $(V(G), \mathscr{L})$ is a convex geometry. For example, if $\mathscr{L}$ is the collection of subsets of $V(G)$ which induce connected subgraphs, then $(V(G), \mathscr{L})$ is a convex geometry if and only if G is a connected block graph [20].

In this paper, we investigate several alignments on graphs and hypergraphs. For each of the alignments that we study, we prove analogues of several classical results, namely, the Minkowski-Krein-Milman theorem, Caratheodory's theorem, Tietze's convexity theorem, and the local convexity of the space (cf. [23], [30]). In addition, our results yield new characterizations of the classes of chordal graphs, strongly chordal graphs, Ptolemaic graphs and totally balanced hypergraphs.

The results in this paper are of a structural nature and, with one exception described below, have no direct applications of which we are aware. Rather, they exploit and highlight the similarities between several classes of graphs and hypergraphs which arise in applications by studying them in the framework of convex geometries. For example, chordal graphs arise in the study of Gaussian elimination with no fill-in [27]. Moreover, they comprise a class of graphs for which simple linear time algorithms exist for several very difficult combinatorial problems, e.g., clique covering, maximum clique, chromatic number, independence number [14], and independent domination number [10]. Strongly chordal graphs are among the very few classes of graphs for which polynomial algorithms are known for the weighted domination, weighted independent domination, connected domination, and cardinality Steiner tree problems [12], [31]. (We note that the results in § 3 were used in [31] to establish a simple polynomial transformation between the connected domination problem and cardinality Steiner tree problem in chordal graphs and strongly chordal graphs. The existence of a polynomial algorithm for the connected domination problem in strongly chordal graphs then followed from the existence of a polynomial algorithm for the cardinality Steiner tree problem in strongly chordal graphs.) Totally balanced hypergraphs find applications in facility location problems [24], in the study of "greedy" linear programming [12], [15], and, as an example of so-called acyclic hypergraphs, in the study of relational database schemes [2].

Some of these results were announced in [21]. Other work related to convex geometries and convexity in graphs can be found in [8], [9], [17], [18]-[22], [25].

2. Preliminaries. If $H=(X, \mathscr{E})$ is a hypergraph and $Y \subseteq X$, then the *neighborhood* of Y, denoted $N(Y)$, is the union of Y and all edges that meet Y. If H is a graph, then $N(Y)$ is the usual closed neighborhood of Y. For every $j>1$, $N^j(Y)=N(N^{j-1}(Y))$, where, of course, $N^1(Y)=N(Y)$.

The distance between two vertices, u and v, will be denoted $d(u, v)$. If H is connected, then the *eccentricity* of a vertex v, denoted $e(v)$, is $\max\{d(u, v): u \in V\}$.

The subhypergraph of H induced (generated) by Y will be denoted $H[Y]$. We say that Y is connected if $H[Y]$ is connected.

If H is a graph and P is a path or cycle in H, then a *chord* of P is an edge of H joining nonconsecutive vertices of P. A chord is *odd* (respectively *even*) if it joins vertices of odd (respectively even) distance from each other in P.

All graphs considered in this paper are assumed to be finite, undirected, simple, and loopless. We refer the reader to [5] ([3]) for graph (hypergraph) theoretic notation and terminology not defined here.

The Caratheodory number of an aligned space $(X, \mathscr{L})$ is the minimum integer k such that, for all $Y \subseteq X$, $\mathscr{L}(Y)$ is the union of the hulls of all subsets A of Y such that $|A| \leqq k$.

3. Chordal graphs and the monophonic alignment. A graph is *chordal* if it contains no cycle of length greater than 3 as an induced subgraph. A vertex is *simplicial* if its neighborhood induces a complete subgraph. The following theorem summarizes several well-known characterizations of chordal graphs.

THEOREM 3.1 [7], [27]. *Let G be a graph. Then the following are equivalent*:

(a) *G is chordal.*

(b) *Every minimal cutset of every induced subgraph of G induces a complete graph.*

(c) *Every induced subgraph of G has a simplicial vertex.*

A set K of vertices of a graph G is *monophonically convex* [21] (m-convex) if K contains every vertex on every chordless path between vertices in K. It is not difficult to see that the family of m-convex sets is closed under intersection. Observe that v is an extreme point of an m-convex set K if and only if v is simplicial in $G[K]$. Hence, the monophonic alignment of a graph G is a convex geometry only if G is chordal. We will show that this necessary condition is also sufficient.

THEOREM 3.2. *In a chordal graph, every nonsimplicial vertex lies on a chordless path between two simplicial vertices.*

We note that this result can be deduced from Dirac's proof that condition (b) of Theorem 3.1 implies condition (c). For completeness, we include a simple direct proof which is a modification of Frank's proof of the same implication [13]. We note that this proof does not assume the existence of simplicial vertices.

Proof of Theorem 3.2. We prove this by induction on the number of vertices in the graph, the base case being trivial.

Let $G=(V, E)$ be a chordal graph on n vertices and suppose that the theorem is valid for every chordal graph on fewer than n vertices. Suppose that v is a nonsimplicial vertex of G. Then v has two nonadjacent neighbors, say u_1 and u_2. Let C be a minimal set of vertices of $V\backslash\{u_1, u_2\}$ which meets all u_1-u_2 paths. Clearly $v\in C$. For $i=1, 2$, let W_i be the vertex set of the component of $G-C$ which contains u_i, and let $G_i=G[W_i\cup C]$. Then C is a minimal cutset of $G[W_1\cup W_2\cup C]$, and so $G[C]$ is a complete graph, by Theorem 3.1. By the inductive hypothesis, either u_i is simplicial in G_i or u_i lies on a chordless path between simplicial vertices of G_i. In either case, G_i has a simplicial vertex, say z_i, in W_i, for $i=1, 2$, since $G[C]$ is a complete graph. Observe that z_i is also simplicial in G. Since C is a minimal cutset in $G[W_1\cup W_2\cup C]$, there is a chordless z_1-v path, say P_1, in $G[W_1\cup\{v\}]$, and a chordless $v-z_2$ path, say P_2, in $G[W_2\cup\{v\}]$. Since C is a cutset, $P_1\cdot P_2$ (the path obtained by concatenating P_1 and P_2) is a chordless path in G joining simplicial vertices and containing v.

The validity of the theorem follows by induction. □

From Theorem 3.2 we obtain an analogue of the Minkowski-Krein-Milman theorem:

COROLLARY 3.3. *If G is chordal, then the monophonic alignment of G is a convex geometry.*

We also obtain a Caratheodory theorem:

COROLLARY 3.4. *The Caratheodory number of the monophonic alignment of a chordal graph is at most* 2.

Proof. Let $G=(V, E)$ be a chordal graph, let S be a subset of V, and let m-conv (S) be the monophonic hull of S. Let $x\in m$-conv (S). If x is simplicial in the subgraph induced by m-conv (S), then $x\in S$, since each extreme point of m-conv (S) is in S, by the definition of the hull of S. Otherwise, x lies on a chordless path between two simplical vertices of the subgraph induced by m-conv (S), by Theorem 3.2, i.e., x is in the monophonic hull of two extreme points of m-conv (S). Thus, x is in the monophonic hull of two vertices in S. □

COROLLARY 3.5. *In a chordal graph $G=(V, E)$, a subset K of vertices is m-convex if and only if there is an ordering $v_1, v_2, \cdots, v_l$ of $V\backslash K$ such that, for each $i=1, 2, \cdots, l$, v_i is simplicial in $G[K\cup\{v_i, v_{i+1}, \cdots, v_l\}]$.*

Proof. This follows immediately from Theorem 1.1, Corollary 3.3, and the relationship between simplicial vertices and extreme points of m-convex sets. □

A basic fact about convexity in R^n is that every ball around every convex set is convex. For m-convexity in chordal graphs we obtain a somewhat stronger result.

THEOREM 3.6. *Suppose $G=(V, E)$ is a chordal graph and K is a connected subset of V. Then $N^j(K)$ is m-convex for every $j\geqq 1$.*

Proof. Notice that $N(K)$ is also connected, and hence it suffices to show that $N(K)$ is m-convex. Suppose, to the contrary, that $N(K)$ is not m-convex. Then there is a chordless path $P=u_0u_1\cdots u_n$, $n>2$, whose intersection with $N(K)$ is exactly $\{u_0, u_n\}$. Hence $\{u_0, u_n\}$ is a minimal cutset in the chordal graph $G[K\cup\{u_0, u_1, \cdots, u_n\}]$, whence $u_0u_n\in E$, by Theorem 3.1, contradicting the fact that P is chordless. □

COROLLARY 3.7. *Suppose G is a connected chordal graph and v is a vertex of G. Then there is a simplicial vertex u such that $d(u, v)=e(v)$.*

Proof. Let $j=e(v)-1$. Then $N^j(v)$ is m-convex. Hence u exists by Corollary 3.5. □

We now obtain an analogue of Tietze's convexity theorem:

THEOREM 3.8. *Suppose that $G=(V, E)$ is a chordal graph and K is a connected subset of V. Then K is m-convex if and only if $N(v)\cap K$ is m-convex, for all $v\in K$.*

Proof. Necessity follows immediately from Theorem 3.6 and the fact that the collection of m-convex sets is closed under intersection. Sufficiency follows from the next proposition. □

PROPOSITION 3.9. *Suppose K is a connected subset of vertices of a graph $G=(V, E)$. Then K is m-convex if $N(v)\cap K$ is m-convex, for all $v\in K$.*

Proof. Suppose K is not m-convex. Then there exists a pair of nonadjacent vertices, u and v, in K and a chordless $u-v$ path $P=uy_1y_2\cdots y_nv$ such that $y_1, y_2, \cdots, y_n\in V\backslash K$. Let $P^*=ux_1x_2\cdots x_kv$ be a shortext $u-v$ path in $G[K]$. If $k=1$, then $N(x_1)\cap K$ is not m-convex. Otherwise, let P' be a shortest $u-x_2$ path in the subgraph induced by $\{u, v, y_1, y_2, \cdots, y_n, x_2, x_3, \cdots, x_k\}$. Then P' contains some vertex in $\{y_1, y_2, \cdots, y_k\}$, by the choice of P^*. Moreover, P' is a chordless path in G. Thus, again, $N(x_1)\cap K$ is not m-convex. □

We note that if the graph G in the above proposition is chordal, then a stronger conclusion can be drawn. Namely, if K is connected but not m-convex, then there exists a pair of vertices u, v in K of distance two from each other in $G[K]$, and a chordless $u-v$ path all of whose interior vertices are in $V\backslash K$. This follows from the proof of the above proposition by observing that if G is chordal, then so is the subgraph H induced by $P\cup P^*$. Thus $N_H(x_1)$ is m-convex in H, and so $P'\subseteq N(x_1)$. But $N(x_1)\cap\{x_3, x_4, \cdots, x_n\}=\emptyset$, by the choice of P^*.

We also note that, for an arbitrary graph $G=(V, E)$ and connected subset K of vertices, if $N(v)\cap K$ is m-convex for all $v\in K$, then K induces a chordal subgraph. Indeed, if C is an induced cycle of length at least four in $G[K]$, then $N(v)\cap K$ is not m-convex, for any $v\in C$.

4. Ptolemaic graphs and the geodesic alignment. A graph is *Ptolemaic* if it is connected and, for every four vertices u, v, w, y the following inequality holds.

$$d(u, v)d(w, y)\leqq d(u, w)d(v, y)+d(v, w)d(u, y).$$

A set K of vertices of a graph G is *geodesically convex* (g-convex) if K contains every vertex on every shortest path between vertices of K. Clearly, the collection of g-convex sets in G is closed under intersection. It is easy to see that v is an extreme point of a g-convex set K if and only if v is simplicial in $G[K]$. Hence the geodesic alignment of G is a convex geometry only if G is chordal. The converse is false. Consider, for example, the chordal graph in Fig. 1, which is called a 3-fan. The vertices v_2 and v_3 clearly do not satisfy the required antiexchange property with respect to the g-convex set $\{v_0, v_1, v_4\}$. Notice that this graph is not Ptolemaic, since $d(v_1, v_3)d(v_2, v_4) > d(v_1, v_2)d(v_3, v_4) + d(v_1, v_4)d(v_2, v_3)$. We will show that the geodesic alignment of G is a convex geometry if and only if G is a disjoint union of Ptolemaic graphs.

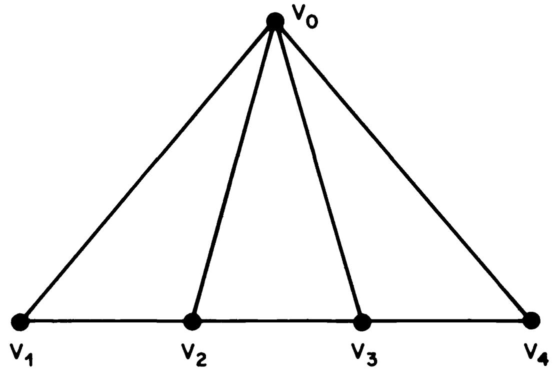


FIG. 1

THEOREM 4.1. *Let $G = (V, E)$ be a graph. Then the following are equivalent*:

(a) *G is a disjoint union of Ptolemaic graphs.*

(b) *G is chordal and every 5-cycle has at least 3 chords.*

(c) *G is chordal and contains no induced 3-fan.*

(d) *G is chordal and all chordless paths are shortest paths.*

(e) *The geodesic alignment of G is a convex geometry.*

(f) *G is chordal and the monophonic and geodesic alignment of G are identical.*

Proof. The equivalence of (a), (c) and (d) is due to Howorka [16], and the equivalence of (b) and (c) is trivial. We will establish 3 implications, namely, (d) implies (e), (e) implies (f), and (f) implies (c).

The fact that condition (d) implies condition (e) follows immediately from Corollary 3.3.

Suppose condition (e) holds. Since an extreme point of a g-convex set K is a simplicial vertex in $G[K]$, G must be chordal, by Theorem 3.1. Moreover, any g-convex set must be m-convex, by Corollary 3.5. Clearly, any m-convex set is g-convex. Thus condition (f) holds.

Finally, suppose G is chordal and contains a 3-fan, say $u_0u_1u_2u_3$ is a chordless path and $vu_i \in E$ for $i = 0, 1, 2, 3$. Observe that $\{u_1, u_2\}$ is contained in the monophonic hull of $\{u_0, u_3\}$. On the other hand, we claim that neither u_1 nor u_2 is in the geodesic hull of $\{u_0, u_3\}$. Let A be the set of vertices lying on shortest $u_0 - u_3$ paths. Observe that u_1, $u_2 \notin A$. Since $d(u_0, u_3) = 2$, $A \backslash \{u_0, u_3\}$ is a minimal cutset in $G[A]$. Hence, $A \backslash \{u_0, u_3\}$ induces a complete subgraph, by Theorem 3.1. It follows that A is g-convex. Hence, condition (f) implies condition (c). □

We note that after this paper was originally submitted for publication, we learned that the equivalence of conditions (c), (d) and (e) was announced by V. P. Soltan in [29]. We have included this proof for two reasons. First, [29] does not contain a proof, and second, our proof is very simple, and highlights the relation between this result and the material in § 3.

5. Totally balanced hypergraphs and the simple path alignment. A *path* in a hypergraph is a sequence $x_1E_1x_2E_2 \cdots x_{n-1}E_{n-1}x_n$ where the x_i's are pairwise distinct vertices, the E_i's are pairwise distinct edges, and x_i, $x_{i+1} \in E_i$, for each $i=1, 2, \cdots, n-1$. A *circuit* is defined in the same way as a path, except $x_1 = x_n$. We say that vertex x is on the path (circuit) $x_1E_1x_2E_2 \cdots x_{n-1}E_{n-1}x_n$ if $x \in \{x_1, x_2, \cdots, x_n\}$. (Thus, some E_i may contain vertices which are not on the path or circuit.) The path (circuit) is *simple* if $E_i \cap \{x_1, x_2, \cdots, x_n\} = \{x_i, x_{i+1}\}$ for $i=1, 2, \cdots, n-1$.

A hypergraph is *totally balanced* if it contains no simple circuit of length greater than 2, i.e., if each partial subhypergraph which is a graph is, in fact, a forest. Observe that every partial subhypergraph of a totally balanced hypergraph is totally balanced. A vertex of a hypergraph is a *nest vertex* (simple vertex [1]) if the edges containing it form a nested family of sets. The following theorem parallels part of Theorem 3.1.

THEOREM 5.1 [1], [6]. *A hypergraph H is totally balanced if and only if every subhypergraph of H has a nest vertex.*

If $H=(X, \mathscr{E})$ is a totally balanced hypergraph and $\mathscr{L}(H)$ is the collection of subsets of X obtainable from X by sequentially deleting nest vertices, then $(X, \mathscr{L}(H))$ is a convex geometry, since the property of being a nest vertex is maintained under taking subhypergraphs. Alan Hoffman posed the problem of finding an internal description of the convex sets in this convex geometry. In light of the similarity in the definitions of chordal graphs and totally balanced hypergraphs, it is reasonable to expect that the description of the associated convex sets will also be similar. Indeed, it is only necessary to replace the word "chordless" by "simple".

In the *simple path alignment* of a hypergraph $H=(X, \mathscr{E})$, a set K of vertices is *s.p. convex* if K contains every vertex on every simple path between vertices of K. It is not difficult to see that the collection of s.p. convex sets in H is closed under intersection. Further, if K is s.p. convex in H, then x is an extreme point of K if and only if x is a nest vertex in $H[K]$. Thus, by Theorem 5.1, the s.p. alignment of H is a convex geometry only if H is totally balanced. We will show that this necessary condition is also sufficient.

Remark 5.2. In a totally balanced hypergraph $H=(X, \mathscr{E})$, a set K of vertices is s.p. convex if and only if $H'=(X, \mathscr{E} \cup \{K\})$ is totally balanced.

LEMMA 5.3. *In a totally balanced hypergraph every minimal cutset is s.p. convex.*

Proof. Let $H=(X, \mathscr{E})$ be a connected hypergraph. Suppose Z is a minimal cutset in H which is not s.p. convex. Then there exists a simple path $u_0E_0u_1E_1 \cdots u_{n-1}E_{n-1}u_n$, $n \geqq 2$, with u_0, $u_n \in Z$ and $u_1, u_2, \cdots, u_{n-1} \notin Z$. Clearly, $u_1, u_2, \cdots, u_{n-1}$ lie together in some component, H_1, of $H-Z$. Let H_2 be another component of $H-Z$, and let H' be the partial hypergraph of H generated by those edges which meet H_2. Since Z is a minimal cutset, there is a simple u_0-u_n path, say $u_0F_1y_2F_2 \cdots y_kF_ku_n$, in H', all of whose interior vertices lie in H_2 (possibly $k=1$). Since Z is a cutset, none of E_0, $E_1, \cdots, E_n$ meets H_2 and none of F_0, $F_1, \cdots, F_k$ meets H_1. Thus, $u_0E_0u_1E_1 \cdots u_{n-1}E_{n-1}u_nF_ky_{k-1} \cdots F_2y_2F_1u_0$ is a simple circuit of length $n+k(\geqq 3)$. Hence, H is not totally balanced. □

The following lemma, which we state without proof, is due independently to Berge [4] and Ryser [28].

LEMMA 5.4. *Suppose $H=(X, \mathscr{E})$ is a hypergraph which does not contain a 3-cycle as a partial subhypergraph. Let $K \subseteq X$. If every pair of vertices in K lies in an edge, then there is some edge containing K.*

The main theorem of this section follows.

THEOREM 5.5. *In a totally balanced hypergraph every nonnest vertex lies on a simple path between nest vertices.*

Proof. We prove this by induction on the number of vertices, the base case being trivial. We may, of course, assume that the hypergraph is connected. Thus, let $H=(X, \mathscr{E})$ be a connected totally balanced hypergraph on n vertices, $n>1$, and assume that the theorem is valid for every totally balanced hypergraph on fewer than n vertices. If $X \in \mathscr{E}$, then each nest vertex (simple path) in the partial hypergraph $(X, \mathscr{E}\backslash\{X\})$ is a nest vertex (simple path) in H. Thus, we may assume that $X \notin \mathscr{E}$. By Lemma 5.4, there exist $x, y \in X$ such that $\{x, y\}$ is contained in no edge, and hence there is a cutset. Let Z be a minimal cutset, and let W_1 and W_2 be the vertex sets of any two components of $H-Z$. By Lemma 5.3, Z is s.p. convex. Let $H_i = H[W_i \cup Z]$ for $i=1, 2$, and let $w \in W_1$. If w is a nest vertex of H_1, then it is a nest vertex of H, because Z is a cutset. Otherwise, w lies on a simple path $P=u_0E_0u_1E_1 \cdots u_{k-1}E_{k-1}u_k$ (say $w=u_i$) between nest vertices of H_1, by the inductive hypothesis. If u_0, $u_k \notin Z$ then u_0 and u_k are nest vertices of H. Otherwise, we may assume that $u_k \in Z$. Let $j=\min\{l\colon u_l \in Z \text{ and } l>i\}$. If $u_p \in Z$ for some $p<i$, then $u_pE_pu_{p+1}E_{p+1} \cdots u_{j-1}E_{j-1}u_j$ is a simple path between vertices of Z which contains $w=u_i$, contradicting the fact that Z is s.p. convex. Thus, $u_0, u_1, \cdots, u_{j-1} \notin Z$. In particular, u_0 is a nest vertex of H. Thus, in any case, there is a nest vertex of H lying in W_1. By symmetry, there is a nest vertex of H, say y, lying in W_2. Let $u_jF_0y_1 \cdots y_tF_ty$ be a simple u_j-y path, all of whose interior vertices lie in W_2. (Such a path exists because Z is a minimal cutset.) Then none of the edges $E_0, E_1, \cdots, E_{j-1}$ meets W_2 and none of the edges $F_0, F_1, \cdots, F_t$ meets W_1, because Z is a cutset. Hence $u_0E_0u_1E_1 \cdots u_{j-1}E_{j-1}u_jF_0y_1F_1 \cdots y_tF_ty$ is a simple path between nest vertices of H which contains w.

Since w was an arbitrary vertex of $V\backslash Z$, it only remains to show that each vertex of Z lies on a simple path between nest vertices. Let $z \in Z$. Let P_1 be a simple path from some nest vertex in W_1 to z, all of whose interior vertices are in W_1, and let P_2 be a simple path from z to some nest vertex in W_2, all of whose interior vertices are in W_2. (P_1 and P_2 exist since Z is a minimal cutset). Then $P_1 \cdot P_2$ is a simple path between nest vertices which contains z.

The validity of the theorem follows by induction. □

We note that in the above proof we did not assume the existence of nest vertices. This proof has some similarity to the proof of the existence of nest vertices appearing in [6] (cf. [26]), as well as obvious similarities to the proof of Theorem 3.2.

COROLLARY 5.6. *The simple path alignment of a hypergraph H is a convex geometry if and only if H is totally balanced.*

Proof. We have already observed the necessity of this condition. Sufficiency follows immediately from Theorem 5.5. □

COROLLARY 5.7. *The Caratheodory number of the simple path alignment of a totally balanced hypergraph is at most* 2.

The proof of this corollary is similar to that of Corollary 3.4, and is omitted.

COROLLARY 5.8. *In a totally balanced hypergraph $H=(X, \mathscr{E})$ a set K of vertices is s.p. convex if and only if there is an ordering $v_1, v_2, \cdots, v_m$ of $V\backslash K$ such that, for $i=1, 2, \cdots, m$, v_i is a nest vertex in the subhypergraph induced by $K \cup \{v_i, v_{i+1}, \cdots, v_m\}$.*

THEOREM 5.9. *Suppose $H=(X, \mathscr{E})$ is a totally balanced hypergraph and K is a connected subset of X. Then $N^j(K)$ is s.p. convex for every $j \geqq 1$.*

Proof. Similar to that of Theorem 3.6. Appeal to Lemma 5.3 rather than Theorem 3.1. □

COROLLARY 5.10. *Suppose v is a vertex of a connected totally balanced hypergraph H. Then H has a nest vertex u such that $d(u, v) = e(v)$.*

Proof. Let $j = e(v) - 1$. Then $N^j(v)$ is s.p. convex. Hence u exists by Corollary 5.8. □

THEOREM 5.11. *Suppose $H = (X, \mathscr{E})$ is a totally balanced hypergraph and K is a connected subset of X. Then K is s.p. convex if and only if $N(v) \cap K$ is s.p. convex, for each $v \in K$.*

Proof. Necessity follows immediately from Theorem 5.9 and the fact that the family of s.p. convex sets with closed under intersection. Sufficiency follows from the next proposition. □

PROPOSITION 5.12. *Suppose K is a connected set of vertices of a hypergraph $H = (V, \mathscr{E})$. Then K is s.p. convex if $N(v) \cap K$ is s.p. convex for all $v \in K$.*

Proof. The proof is similar to that of Proposition 3.9, except:

(i) Edges, as well as vertices, must be specified.

(ii) The vertices u and v may lie together in some edge, in which case $N(v) \cap K$ is not s.p. convex. □

As with Proposition 3.9, a stronger conclusion can be drawn if H is totally balanced. Namely, if K is connected but not m-convex, then there exist two vertices of K of distance at most two from each other in K which are joined by a simple path of length at least two, all of those interior vertices are in $V \backslash K$. The proof is again similar to the corresponding proof for chordal graphs.

Unlike the situation for chordal graphs, if H is not totally balanced, then K need not induce a totally balanced subhypergraph if K is connected and $N(v) \cap K$ is s.p. convex for all $v \in K$. For example, let H be the complete graph on three vertices, and let K consist of all three vertices.

6. Strongly chordal graphs and the strong alignment. A graph is *strongly chordal* [11] if it is chordal and, in addition, every even cycle of length at least 6 has an odd chord.

A vertex of a graph is *simple* if the neighborhoods of its neighbors form a nested family of sets. Notice that a simple vertex must be simplicial, but not conversely. The *neighborhood hypergraph*, $\mathscr{N}(G)$, of a graph $G = (V, E)$ is the hypergraph $(V, \{N(v): v \in V\})$. Observe that v is simple in G if and only if it is a nest vertex of $\mathscr{N}(G)$. The following theorem summarizes several characeterizations of strongly chordal graphs.

THEOREM 6.1 [11]. *Let G be a graph. Then the following are equivalent*:

(a) *G is strongly chordal.*

(b) *Every induced subgraph of G has a simple vertex.*

(c) *$\mathscr{N}(G)$ is totally balanced.*

Given a strongly chordal graph $G = (V, E)$, the collection of subsets of V obtainable from V by sequentially deleting simple vertices defines a convex geometry on V, since the property of being a simple vertex is maintained under taking induced subgraphs. In light of the results in the previous sections, one might suspect that this convex geometry can be defined in terms of closure under certain paths. As we show below, this is indeed the case. In view of Corollary 5.6 and Theorem 6.1, one might suspect that this convex geometry is nothing other than the simple path alignment of $\mathscr{N}(G)$. The latter suspicion is incorrect. For example, in the graph depicted in Fig. 2, $\{1, 7\}$ can be obtained by deleting 2, 6, 4, 3, and then 5, each of which is simple when it is deleted, and yet $\{1, 7\}$ is not s.p. convex in the neighborhood hypergraph.

Let $G = (V, E)$ be a graph. We say that a path $P = u_0u_1u_2 \cdots u_n$ is *even-chorded* if it has no odd chords and, in addition, neither u_0 nor u_n lies on a chord of P. A set

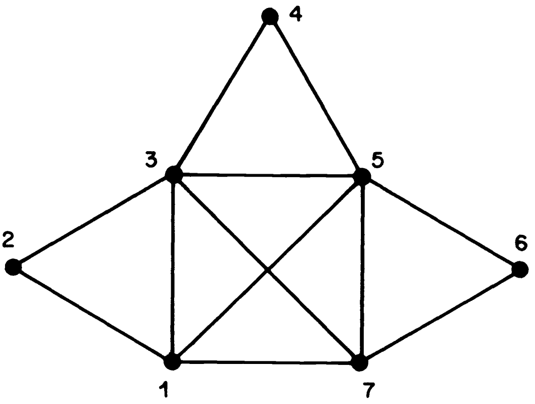


FIG. 2

K of vertices is *strongly convex* (s-convex) if K contains every vertex on every even-chorded path whose endpoints are in K. It is not difficult to see that the collection of s-convex sets is an alignment of G, which we call the *strong alignment.* Since chordless paths are even-chorded, an extreme point of an s-convex set K must be simplicial in $G[K]$. Hence, the strong alignment of G is not a convex geometry unless G is chordal. In the case that G is chordal, it is not difficult to see that v is an extreme point of an s-convex set K if and only if v is simple in $G[K]$. Thus the strong alignment of G is a convex geometry only if G is strongly chordal. We will show that this necessary condition is also sufficient.

LEMMA 6.2. *Let $P = u_0u_1u_2 \cdots u_n$ be a path in a chordal graph. Then P has an odd chord if and only if P has a pair of consecutive chords, i.e., a pair of chords u_iu_j and u_iu_{j+1}.*

Proof. If u_iu_j and u_iu_{j+1} are chords of P, then one of them is an odd chord.

Suppose P has an odd chord but has no pair of consecutive chords. Choose an odd chord u_iu_j with $i<j$ and $j-i$ as small as possible. Consider the cycle $C = u_iu_{i+1} \cdots u_ju_i$. By assumption, u_iu_{j-1} is not a chord, and hence u_j lies on some chord of C. Let u_ku_j be that chord of C which minimizes k. By assumption, $k \neq i+1$. Thus, u_iu_k is also a chord of C. Hence, either u_iu_k or u_ku_j is an odd chord of P, contradicting the choice of u_iu_j. □

THEOREM 6.3. *In a strongly chordal graph, every nonsimple vertex lies on an even-chorded path between simple vertices.*

Proof. Let $G = (V, E)$ be a strongly chordal graph and suppose that $v \in V$ is not simple. Then v is not a nest vertex in $\mathcal{N}(G)$. Thus, in $\mathcal{N}(G)$, v lies on a simple path $P = v_0N(v_1)v_2N(v_3) \cdots N(v_{2n-1})v_{2n}$ between nest vertices (say $v = v_{2t}$), by Theorems 5.5 and 6.1. Now, v_0 and v_{2n} are simple in G. Notice that $v_i = v_{i+1}$ is possible, but $v_i \neq v_j$ if $|i-j| > 1$. Also, if we consider the sequence $v_0v_1v_2 \cdots v_{2n}$ without repetition, we obtain a path $P^* = x_0x_1 \cdots x_k$ in G.

We claim that P^* has no odd chords. By Lemma 6.2, it suffices to show that P^* has no consecutive chords. Suppose, to the contrary, that x_ix_j and x_ix_{j+1} are both chords of P^*. If $x_i = v_{2k}$, for some k, then either x_j or x_{j+1} is v_{2l+1}, for some $l \notin \{k, k-1\}$. Thus $v_{2k} \in N(v_{2l+1})$, contradicting the fact that P is a simple path. A similar contradiction is obtained under the assumption that $x_i = v_{2k+1}$, for some k, proving the claim.

If x_0 and x_k lie on no chords of P^*, then P^* is even-chorded, and we are done. Thus, suppose x_0 lies on a chord, say $x_0x_i \in E$. Then $x_i = v_{i'}$ for some $i' \geqq i \geqq 2$. Since P is a simple path in $\mathcal{N}(G)$ and $v_0 = x_0 \in N(v_{i'})$, i' must be even. Since v_0 is a simple vertex and $v_{i'}$, $v_1 \in N(v_0)$, we have $v_{i'} \in N(v_1)$. Since P is a simple path and i' is even, we have $i'=2$, and so $i=2$. In that case, $x_0x_2x_3\cdots x_n$ is a path containing $v=v_{2t}$ (since we only deleted v_1) which has no odd chords and in which x_0 lies on no chord. By a similar argument, if x_k lies on a chord of P, then that chord is $x_{k-2}x_k$. Thus, by deleting x_1 and/or x_{k-1}, if necessary, we obtain an even-chorded path between simple vertices of G which contains v. □

COROLLARY 6.4. *The strong alignment of a graph G is a convex geometry if and only if G is strongly chordal.*

COROLLARY 6.5. *The Caratheodory number of the strong alignment of a strongly chordal graph G is at most* 2.

COROLLARY 6.6. *In a strongly chordal graph $G=(V,E)$ a set K of vertices is s-convex if and only if there is an ordering $v_1, v_2, \cdots, v_m$ of $V\backslash K$ such that, for each $i=1,2,\cdots,m$, v_i is simple in $G[K\cup\{v_i, v_{i+1},\cdots,v_m\}]$.*

In light of this corollary and the example at the beginning of this section, we find that a set which is s-convex in the strongly chordal graph G need not be s.p. convex in $\mathcal{N}(G)$. On the other hand, we have the following result.

PROPOSITION 6.7. *If $G=(V,E)$ is a strongly chordal graph and $K\subseteq V$ is s.p. convex in $\mathcal{N}(G)$, then K is s-convex in G.*

Proof. Suppose K is s.p. convex in $\mathcal{N}(G)$. Then, by Theorem 6.1 and Corollary 5.8, there is an ordering $v_1, v_2, \cdots, v_m$ of $V\backslash K$ such that, for $i=1,2,\cdots,m$, v_i is a nest vertex in the subhypergraph of $\mathcal{N}(G)$ induced by $K\cup\{v_i, v_{i+1},\cdots,v_m\}$. It follows that v_i is simple in $G[K\cup\{v_i, v_{i+1},\cdots,v_m\}]$, for each i. Thus K is s-convex in G, by Corollary 6.6. □

THEOREM 6.8. *Suppose $G=(V,E)$ is strongly chordal and K is a connected subset of V. Then $N^j(K)$ is s-convex for all $j\geqq 2$. Moreover, $N(v)$ is s-convex for all $v\in V$.*

Proof. Observe that $N(K)$ is also connected. Thus, to prove the first claim, it suffices to show that $N^2(K)$ is s-convex. Observe that K induces a connected subhypergraph of $\mathcal{N}(G)$ and that $N^2(K)$ is precisely the neighborhood of K in $\mathcal{N}(G)$. Thus $N^2(K)$ is s.p. convex in $\mathcal{N}(G)$, by Theorems 5.9 and 6.1. Hence, $N^2(K)$ is s-convex in G, by Proposition 6.7.

Since $N(v)$ is an edge of $\mathcal{N}(G)$, it is s.p. convex in $\mathcal{N}(G)$, by Remark 5.2 and Theorem 6.1. Hence $N(v)$ is s-convex in G, by Proposition 6.7. □

We note that $N(K)$ may not be s-convex even if K is s-convex. For example, in the graph depicted in Fig. 2, $N(\{1,7\})$ is not s-convex.

COROLLARY 6.9. *Suppose v is a vertex of a connected strongly chordal graph G. Then G has a simple vertex u such that $d(u,v)=e(v)$.*

LEMMA 6.10. *Suppose $P=u_0u_1u_2\cdots u_n$ is an even-chorded path in a strongly chordal graph G, and u_iu_j is a chord of P, with $i<j$. Then $\{u_i, u_{i+2}, u_{i+4},\cdots,u_j\}$ induces a complete graph and $\{u_{i+1}, u_{i+3},\cdots,u_{j-1}\}$ is independent.*

Proof. We prove this by induction on $j-i$, the case $j-i=2$ being trivial. Suppose that the lemma holds for $j'-i'<k$ and that $j-i=k$. Consider the cycle $C=u_iu_{i+1}\cdots u_ju_i$. Since P is even-chorded, u_iu_{j-1} is not a chord of C, and hence u_j lies on some chord of C. Let u_lu_j be that chord of C which minimizes l. Then $l=i+2$, for otherwise $u_iu_{i+1}\cdots u_lu_ju_i$ is an even cycle of length at least 6 with no odd chords, contradicting the fact that G is strongly chordal. By symmetry, u_iu_{j-2} is also a chord. By the inductive hypothesis, $\{u_i, u_{i+2},\cdots,u_{j-2}\}$ and $\{u_{i+2}, u_{i+4},\cdots,u_j\}$ induce complete graphs and $\{u_{i+1}, u_{i+3},\cdots,u_{j-3}\}$ and $\{u_{i+3}, u_{i+5},\cdots,u_{j-1}\}$ are independent. Since $u_iu_j\in E(G)$, it only remains to show that $u_{i+1}u_{j-1}\notin E(G)$. If $u_{i+1}u_{j-1}\in E(G)$, then either $u_iu_{i+1}u_{j-1}u_ju_i$ is an induced 4-cycle of G, or P has an odd chord, contradicting the hypothesis.

The validity of the lemma follows by induction. □

THEOREM 6.11. *Suppose $G = (V, E)$ is strongly chordal and K is a connected subset of V. Then K is s-convex if and only if $N^2(v) \cap K$ is s-convex, for all $v \in K$.*

Proof. Necessity follows immediately from Theorem 6.8 and the fact that the family of s-convex sets is closed under intersection.

Suppose K is not s-convex. If K is not m-convex, then it contains a vertex v such that $N(v) \cap K$ is not m-convex, by Theorem 3.8. Since $N(v)$ is m-convex, and the family of m-convex sets is closed under intersection it follows that $N^2(v) \cap K$ is not m-convex. Hence, $N^2(v) \cap K$ is not s-convex. Thus, suppose K is m-convex. Let $P = u_0 u_1 \cdots u_n$ be a shortest even-chorded path between vertices of K which meets $V \backslash K$. We will show that $u_1 \in K$ and $u_1 u_{n-1} \in E$, and hence that $N^2(u_1) \cap K$ is not s-convex.

Observe that u_1 lies on a chordless path between u_0 and u_n. Hence, $u_1 \in K$ since K is m-convex. Let $j = \max \{i: u_1 u_i \in E\}$. If $j = n-1$ we are done. So, suppose $j < n-1$. By the choice of P, the even-chorded path $u_1 u_j u_{j+1} \cdots u_n$ does not meet $V \backslash K$. Thus, $j > 2$, since P meets $V \backslash K$. Let $l = \min\{i: u_i u_{j+1} \in E\}$. Then, $2 \leqq l \leqq j$. Again, by the choice of P, the even-chorded path $u_0 u_1 \cdots u_l u_{j+1}$ does not meet $V \backslash K$. In particular, $u_2 \in K$. Consequently, u_2 lies on some chord of P, for otherwise $u_2 u_3 \cdots u_n$ would contradict the choice of P. Thus, $u_2 u_4 \in E$, by Lemma 6.10. On the other hand, $\{u_2, u_4, u_6, \cdots, u_{j-1}\}$ is independent, again by Lemma 6.10, since $u_1 u_j \in E$. Hence $j = 3$, whence P does not meet $V \backslash K$, contradicting the choice of P. □

We note that, by Theorem 6.8, if $G = (V, E)$ is strongly chordal and $K \subseteq V$ is s-convex then $K \cap N(v)$ is s-convex for each $v \in K$. However, the converse is false. The set $\{1, 2, 3, 5, 6, 7\}$ in the graph in Fig. 2 is a counterexample.

We also note that, as with Theorems 3.8 and 5.11, the sufficiency of Theorem 6.11 holds for all graphs. However, in this case, while the "form" of Teitze's theorem is true, the "substance" is not. To be precise, it is possible for a graph to contain a connected set of vertices K which is not s-convex, but which has the property that the hull of $N^2(v) \cap K$ lies entirely in K, for all $v \in K$. (In the language of Valentine [30], a connected, strongly locally convex set is convex, but a connected, weakly locally convex set need not be convex.) For example, consider the graph G in Fig. 3. Let K consist of all vertices except v_8. It is straightforward to verify that the only even chorded path between vertices of K containing v_8 is $v_1 v_2 v_3 v_4 v_5 \cdots v_{15}$. Since $d(v_1, v_{15}) = 5$ and v_1 and v_5 have degree 1, we deduce that the hull of $N^2(v) \cap K$ is contained in K for all $v \in K$. On the other hand, $N^2(v_1) \cap K$ is not s-convex, since, e.g., $v_3 v_5 v_6$ is a chordless path between vertices of $N^2(v_1) \cap K$ which is not contained in $N^2(v_1) \cap K$.

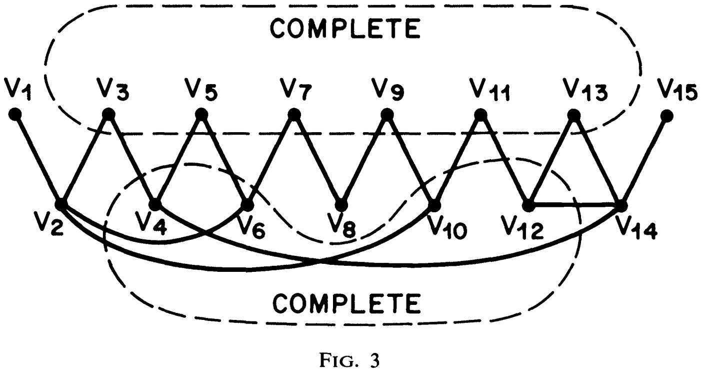


FIG. 3

Acknowledgments. We would like to express our gratitude to Dr. Richard Anstee, with whom we had numerous conversations concerning this paper, and Dr. Alan J.

Hoffman, who suggested the problem of finding an internal description of the convex sets of a totally balanced hypergraph.

REFERENCES

[1] R. P. ANSTEE AND M. FARBER, *Characterizations of totally balanced matrices*, J. Algorithms, 5 (1984), pp. 215-230.

[2] C. BERRI, R. FAGIN, D. MAIER AND M. YANNAKAKIS, *On the desirability of acyclic database schemes*, J. Assoc. Comput. Mach., 30 (1983), pp. 479-513.

[3] C. BERGE, *Graphs and Hypergraphs*, North-Holland, Amsterdam 1983.

[4] ———, *Balanced matrices*, Math. Prog., 2 (1972), pp. 19-31.

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

[6] A. E. BROUWER AND A. KOLEN, *A super-balanced hypergraph has a nest point*, Report ZW 146, Mathematisch Centrum, Amsterdam, 1980.

[7] G. A. DIRAC, *On rigid circuit graphs*, Abh. Math. Seminar Univ. Hamburg, 25 (1961), pp. 71-76.

[8] P. DUCHET AND H. MEYNIEL, *Ensembles convèxes dans les graphes* I, *théorèmes de Helly et de Radon pour graphes et surfaces*, European J. Combin., 4 (1983), pp. 127-132.

[9] P. EDELMAN, *Meet-distributive lattices and the anti-exchange closure*, Algebra Universalis, 10 (1980), pp. 290-299.

[10] M. FARBER, *Independent domination in chordal graphs*, Oper. Res. Lett., 1 (1982), pp. 134-138.

[11] ———, *Characterizations of strongly chordal graphs*, Disc. Math., 43 (1983), pp. 173-189.

[12] ———, *Domination, independent domination, and duality in strongly chordal graphs*, D.A.M., 7 (1984), pp. 115-130.

[13] A. FRANK, *Some polynomial algorithms for certain graphs and hypergraphs*, Congressus Numerantium, XV (1976), pp. 211-226.

[14] 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.

[15] A. J. HOFFMAN, A. W. J. KOLEN AND M. SAKAROVITCH, *Totally-balanced and greedy matrices*, this Journal, 6 (1985), pp. 721-730.

[16] E. HOWORKA, *A characterization of Ptolemaic graphs*, J. Graph Theory, 5 (1981), pp. 323-331.

[17] R. E. JAMISON, *A development of axiomatic convexity*, Tech. Report #481, Clemson University, Clemson, SC, 1970.

[18] ———, *Tietze's convexity theorem for semilattices and lattices*, Semigroup Forum, 15 (1978), pp. 357-373.

[19] ———, *Copoints in antimatroids*, Congressus Numerantium, 29 (1980), pp. 535-544.

[20] ———, *Convexity and block graphs*, Congressus Numerantium, 33 (1981), pp. 129-142.

[21] ———, *A perspective on abstract convexity: classifying alignments by varieties*, in Convexity and Related Combinatorial Geometry, D. C. Kay and M. Breen, eds., Marcel Dekker, New York, 1982.

[22] R. E. JAMISON AND R. NOWAKOWSKI, *A Helly theorem for convexity in graphs*, Tech. Report #419, Clemson Univ., Clemson, SC, 1983.

[23] J. L. KELLEY, I. NAMIOKA ET AL., *Linear Topological Spaces*, Springer-Verlag, New York, 1976.

[24] A. KOLEN, *Solving covering problems and uncapacitated plant location problem on trees*, European J. Oper. Res., 12 (1983), pp. 266-278.

[25] B. KORTE AND L. LOVASZ, *Shelling structures, convexity and a happy end*, Report #83274—OR, Institute of Operations Research, Univ. Bonn, 1983.

[26] A. LUBIW, *Γ-free matrices*, Masters thesis, Univ. Waterloo, Waterloo, Ontario, October, 1982.

[27] D. ROSE, *Triangulated graphs and the elimination process*, J. Math. Anal. Appl., 32 (1970), pp. 579-609.

[28] H. J. RYSER, *Combinatorial configurations*, SIAM J. Appl. Math., 17 (1969), pp. 593-602.

[29] V. P. SOLTAN, *d-convexity in graphs*, Soviet Math. Dokl., 28 (1983), pp. 419-421.

[30] F. A. VALENTINE, *Convex Sets*, McGraw-Hill, San Francisco, 1964.

[31] K. WHITE, M. FARBER AND W. PULLEYBLANK, *Steiner trees, connected domination and strongly chordal graphs*, Networks, 15 (1985), pp. 109-124.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

RELATIONSHIPS BETWEEN l^1 METRICS ON RANKINGS: THE CASE OF TIES*

WADE D. COOK† AND MOSHE KRESS‡

Abstract. This paper provides a direct extension to the space of weak orderings of previous work [SIAM J. Appl. Math., 44 (1984), pp. 209-220] on linear rankings. We extend the model of Blin to this larger space and show how to develop position and object based consensus measures which are equivalent to the Cook and Seiford [Management Sci., 24 (1978), pp. 1721-1732] and Kemeny and Snell [in *Mathematics Models in the Social Sciences*, Ginn, New York, 1962, pp. 9-23] models respectively. One result of this development is a characterization of the *KS* model as an integer quadratic programming problem.

Key words. rankings, consensus, preferences, distance measures

1. Introduction. In a recent paper [6] it was shown that for the restricted case of linear orderings, the ranking model of Blin [3] can be used as a *core* from which to generate representations which will allow for degree of disagreement between voter preferences. It was demonstrated that if the Blin model is modified to take into consideration the number of pairs of *preferences* where voters disagree, the vector model of Cook and Seiford [7] (CS) results. Consideration of pairs of *objects* where disagreement is present leads to the representation of Kemeny and Snell [10] (KS). The latter is shown to be a quadratic assignment problem.

In reference to the above it must be pointed out that many different types of ranking problems have been examined in the literature, some of which take the form of quadratic assignment models. Blin and Whinston [4], for example, have shown that the simple majority rule consensus can be derived by solving such a model. Huber and Schultz [9] demonstrate that the maximum likelihood paired-comparison ranking of a set of objects can be determined through the use of such a quadratic model. Similar problems have been looked at by Slater [11], de Cani [8] and others. Barthelemy and Monjardet [2] survey a wide range of consensus ranking literature and discuss mathematical programming model aspects. This literature, however, relates to the probems discussed in [6] and herein only in the sense that both examine ranking (and consensus to some extent). The problems which we deal with herein, in particular that involving the representation of ties in the KS framework through the use of quadratic programming models, have not (to the best of the authors' knowledge) been discussed elsewhere.

In the present paper the results of [6] are extended to the space of weak orderings (linear and tied preferences). Specifically, we show that by beginning with a generalized version of the Blin model, the Cook and Seiford vector model and the Kemeny and Snell matrix model can be derived as natural extensions of that model. Section 2 presents the generalization of the Blin Model to the space of weak orderings. Sections 3 and 4 develop position and object based measures of voter disagreement which are shown to be equivalent to the CS and KS models respectively.

2. A binary preference representation for weak orderings. In [6] it was shown that the *simple binary* model of Blin [3] can act as a basis or core for other more sophisticated representations of voter preference. Blin represents preferences via a binary permutation

* Received by the editors March 30, 1984, and in revised form September 1, 1985. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada under grant A8966.

† Faculty of Administrative Studies, York University, Toronto, Ontario, Canada M3J 2R6.

‡ CEMA, P.O. Box 2250, Haifa, Israel.

matrix $\alpha = (\alpha_{ij})$ where $\alpha_{ij} = 1$ if object i has rank j, and is 0 otherwise. He then defines a distance measure on the space of all linear orderings by $d_B(A, B) = \sum_{i,j} |\alpha_{ij} - \beta_{ij}|$, where α and β are the matrix representations of orderings A and B. By introducing *two different* measures of degree of disagreement into the Blin structure (position-based and object-based disagreement) we arrive at the Cook and Seiford distance [7] and the Kemeny and Snell distance [10] respectively.

In order to extend the results of [6] to the larger space of weak orderings, it is necessary to modify the Blin structure. In the case where A is a weak ordering we define α_{ij} by

$$(2.1)\qquad \alpha_{ij} = \begin{cases} 1 & \text{if object } i \text{ has rank } j,\ i = 1, 2, \cdots, n,\ j = 1, 1.5, 2, 2.5, \cdots, n, \\ 0 & \text{otherwise.} \end{cases}$$

The rationale for using the values $j = 1, 1.5, \cdots$ as rank positions has been discussed elsewhere (see e.g. [7]). For example, the α-matrix for the ranking in which object b is ranked in first place and a and c are tied in second place (i.e. $A = (\genfrac{}{}{0pt}{}{b}{a,\, c})$) is

$$\begin{array}{r|ccccc} & & & \text{Rank} & & \\ & 1 & 1.5 & 2 & 2.5 & 3 \\ \hline a & 0 & 0 & 0 & 1 & 0 \\ \alpha = b & 1 & 0 & 0 & 0 & 0 \\ c & 0 & 0 & 0 & 1 & 0 \end{array}$$

For two rankings A, B (with matrix representations α and β), we define the *extended* Blin distance

$$(2.2)\qquad d_B(A, B) = \sum_{i=1}^{n} \sum_{j=2}^{2n} |\alpha_{i,j/2} - \beta_{i,j/2}|.$$

For a set of weak orderings $\{A_l\}_{l=1}^{m}$, the consensus problem (following along in the same line as in Blin [3]) is to find a ranking $X \equiv (x_{ij})$ which minimizes

$$\sum_{l=1}^{m} \sum_{i=1}^{n} \sum_{j=2}^{2n} |\alpha^{l}_{i,j/2} - x_{i,j/2}|.$$

Defining $\phi_{ij} = \sum_{l=1}^{m} (1 - \alpha^{l}_{ij})$, it is seen that this problem is equivalent to the *minimal disagreement* problem

$$(2.3)\qquad \underset{X=(x_{ij})}{\text{Minimize}} \quad \sum_{i=1}^{n} \sum_{j=2}^{2n} \phi_{i,j/2} x_{i,j/2}$$

$$(2.4)\qquad \text{subject to} \quad \sum_{j=2}^{2n} x_{i,j/2} = 1, \qquad i = 1, 2, \cdots, n,$$

$$(2.5)\qquad \sum_{i=1}^{n} x_{i,j/2} = D_{j/2}, \qquad j = 2, 3, \cdots, 2n$$

$$(2.6)\qquad x_{i,j/2} \in \{0, 1\},$$

where the $(2n-1)$-dimensional vector of $D_{j/2}$ must constitute a valid ranking.

Armstrong et al. [1] have shown that if one defines

$$(2.7)\qquad D_{j/2} = \sum_{s=1}^{j-1} Z_{j-s,s}$$

where

$$Z_{rs}=\begin{cases}1 & \text{if ranks } r \text{ and } s \text{ are combined to yield a tie at position } (r+s)/2,\\ 0 & \text{otherwise,}\end{cases}$$

then $\{D_{j/2}\}_{j=2}^{2n}$ will constitute a ranking for any $\{Z_{rs}\}$ satisfying the conditions

(2.8) $$Z_{j-s,s}-Z_{j-s-1,s+1}\geqq 0, \qquad s\geqq j-s,$$

(2.9) $$Z_{rs}-Z_{sr}=0,$$

(2.10) $$\sum_{r=1}^{n} Z_{rs}=1=\sum_{s=1}^{n} Z_{rs},$$

(2.11) $$Z_{rs}\in\{0,1\}.$$

The extended Blin consensus problem is, therefore, a 0–1 integer programming problem (2.3)–(2.11).

In the following sections we begin with this binary disagreement model, and show how to construct position and object based models which account for *degree* of disagreement. This development (1) serves to provide a direct extension of previous results in [6] to the larger space of weak orderings, and (2) provides a mechanism for deriving a mathematical programming formulation of the KS model.

3. Position based distance. Following the terminology of [6], define the *forward indicator* vectors

(3.1) $$(P^{+}(j/2))_k=\begin{cases}1 & \text{if object } k \text{ is ranked lower (worse) than rank } j/2,\\ & j=2,3,\cdots,2n,\\ 0 & \text{otherwise,}\end{cases}$$

and *backward indicator* vectors

(3.2) $$(P^{-}(j/2))_k=\begin{cases}1 & \text{if object } k \text{ is ranked higher (better) than rank } j/2,\\ & j=2,3,\cdots,2n,\\ 0 & \text{otherwise,}\end{cases}$$

Property 3.1.

$$P^{+}(j/2)=\sum_{t=2j+1}^{2n}\alpha_{\cdot,t/2}, \qquad P^{-}(j/2)=\sum_{t=2}^{2j-1}\alpha_{\cdot,t/2}.$$

Property 3.2 (orthogonality).

$$\langle P^{+}(j/2),P^{-}(j/2)\rangle=0.$$

In [6] the position-based distance between two linear rankings A, B was defined to be

$$d_p(A,B)=n(n-1)-\sum_{j=1}^{n}\left[\langle P_A^{+}(j),P_B^{+}(j)\rangle+\langle P_A^{-}(j),P_B^{-}(j)\rangle\right]$$

which is the number of situations in which for a rank position j and object k, that the object does not lie on the same side (above or below) of j in both A and B. It was shown that the maximum agreement between A and B when they are identical, is equal to $n(n-1)$.

THEOREM 3.1. *If A and B are identical*

$$\sum_{j=2}^{2n} \langle P_A^+(j/2), P_B^+(j/2)\rangle = \sum_{j=2}^{2n} \langle P_A^-(j/2), P_B^-(j/2)\rangle = n(n-1).$$

Proof. See Appendix.

It is noted that the maximum achievable agreement is twice what it is for the case of the space of strict linear orderings.

DEFINITION 3.1. The *position based* distance $d_p(A, B)$ in the space of weak orderings is given by

$$d_p(A, B) = 2n(n-1) - \sum_{j=2}^{2n} [\langle P_A^+(j/2), P_B^+(j/2)\rangle + \langle P_A^-(j/2), P_B^-(j/2)\rangle].$$

The proof of the following theorem is similar to that given in [6] for strict linear orderings, and is, therefore, omitted.

THEOREM 3.2. *The position based distance between any two weak orderings A and B is twice the* CS *distance, i.e.,*

$$d_p(A, B) = 2d_{CS}(A, B).$$

Consensus formation.

DEFINITION 3.2. The *consensus ranking X* for the position based distance is that ranking (extended Blin matrix (x_{ij})) which minimizes

$$\sum_{l=1}^{m} d_p(A^l, X) = \sum_{l=1}^{m}\left[2n(n-1) - \sum_{j=2}^{2n} [\langle P_l^+(j/2), X^+(j/2)\rangle + \langle P_l^-(j/2), X^-(j/2)\rangle]\right].$$

This problem, and hence the CS consensus problem, is equivalent to the 0-1 linear integer programming model

$$\text{Maximize} \quad \sum_{l=1}^{m}\sum_{i=1}^{n}\sum_{j=2}^{2n}\left[\left(\sum_{t=2}^{j-1}\alpha_{i,t/2}^{l}\right)\left(\sum_{t=2}^{j-1}x_{i,t/2}\right)+\left(\sum_{t=j+1}^{2n}\alpha_{i,t/2}^{l}\right)\left(\sum_{t=j+1}^{2n}x_{i,t/2}\right)\right]$$

subject to constraints (2.4)-(2.11).

Hence, we have shown that building on an extended version of Blin's model for weak orderings and accounting for agreement as to pairs of ranks, one can derive a consensus model equivalent to that of Cook and Seiford [7].

We now examine object based agreement.

4. Object based distance. One of the principal results established in [6] was the characterization of the KS consensus model as a quadratic assignment problem. As indicated in the introduction, a number of other ranking or consensus models (e.g. the majority rule consensus and the maximum likelihood paired-comparison ranking) can also be viewed in terms of such quadratic models. In order to derive the appropriate model in the space of weak orderings it is necessary to deal properly with tied preferences. Let us define

$$Q_k^+(i) = \begin{cases} 1 & \text{if object } i \text{ is preferred to object } k, \\ 0 & \text{otherwise,} \end{cases}$$

$$Q_k^-(i) = \begin{cases} 1 & \text{if object } k \text{ is preferred to object } i, \\ 0 & \text{otherwise,} \end{cases}$$

$$Q_k^0(i) = \begin{cases} 1 & \text{if objects } i \text{ and } k \text{ are tied,} \\ 0 & \text{otherwise.} \end{cases}$$

Letting j_i be the rank of i (i.e. $\alpha_{i,j_i}=1$), we have
Property 4.1.

$$Q^+(i)=\sum_{t=2j_i+1}^{2n}\alpha_{\cdot,t/2},\quad Q^-(i)=\sum_{t=2}^{2j_i-1}\alpha_{\cdot,t/2},\quad Q^0(i)=\alpha_{\cdot,j_i}.$$

In order to proceed from the Blin model to that of Kemeny and Snell, we define object based distance as follows:

DEFINITION 4.1. The *object based* distance $d_Q(A, B)$ in the space of weak orderings is given by

$$d_Q(A, B)=n(n-1)-\sum_{i=1}^{n}[\langle Q_A^+(i), Q_B^+(i)\rangle+\langle Q_A^-(i), Q_B^-(i)\rangle+\tfrac{1}{2}\langle Q_A^0(i), Q_B^0(i)\rangle].$$

The proof of the following theorem is similar to that given in [6] for linear orderings, and is therefore omitted.

THEOREM 4.1. *In the space of weak orderings* $d_Q(A, B)=d_{KS}(A, B)$, *where* d_{KS} *is the Kemeny-Snell distance.*

From this theorem it is clear why it is necessary to use a multiplier of $\frac{1}{2}$ in the d_Q term corresponding to ties in A and B.

Consensus formation.

DEFINITION 4.2. The *consensus ranking X* for the object based distance is that ranking which minimizes

$$\sum_{l=1}^{m}d(A^l, X)=\sum_{l=1}^{m}\left[n(n-1)-\sum_{i=1}^{n}[\langle Q_{A^l}^+(i), X^+(i)\rangle+\langle Q_{A^l}^-(i), X^-(i)\rangle+\tfrac{1}{2}\langle Q_{A^l}^0(i), X_i^0\rangle]\right].$$

This problem, and hence the KS problem, can be written in the form (from Property 4.1)

$$\text{Maximize}\quad \sum_{l=1}^{m}\sum_{i=1}^{n}\sum_{k=1}^{n}\left[\sum_{j=2}^{2n}\left\{x_{i,j/2}(\alpha_{i,k}^+(l)\sum_{t=j+1}^{2n}x_{k,t/2}\right.\right.$$
$$\left.\left.+\alpha_{i,k}^-(l)\sum_{t=2}^{j-1}x_{k,t/2}+\alpha_{k,j_i(l)/2}^l\cdot x_{k,j_i(l)/2}\right\}\right],$$

subject to constraints (2.4)-(2.11) where $j=j_i(l)$ is such that $\alpha_{i,j/2}^l=1$,

$$\alpha_{i,k}^+(l)=\sum_{t=j_i(l)+1}^{2n}\alpha_{k,t/2}^l\quad\text{and}\quad\alpha_{i,k}^-(l)=\sum_{t=2}^{j_i(l)-1}\alpha_{k,t/2}^l.$$

This integer quadratic programming problem is a generalized version of the quadratic assignment problem of [6] for linear orderings. The constraint set has been modified (expanded) to permit more than a single object to be ranked at a given position, and there is a linear term in the objective function to account for tied preferences.

5. Summary and conclusions. This paper has directly extended the results of previous work [6] to the space of all weak orderings, and has provided a framework for comparing three well-known consensus methods in that space. The development herein has also provided a vehicle for characterizing the mathematical programming structure of the KS consensus model.

Appendix.

Proof of Theorem 3.1. Let A_n and B_n be weak orderings with n objects. If $A_n \equiv B_n$, then

$$\sum_{j=2}^{2n} \langle P^+_{A_n}(j/2), P^+_{B_n}(j/2)\rangle = \sum_{j=2}^{2n} \|P^+_{A_n}(j/2)\|^2.$$

We proceed by induction. For $n=1$, the result is trivial. For $n=2$,

$$\sum_{j=2}^{2n} \|P^+_{A_n}(j/2)\|^2 = 0+0+2 = n(n-1)$$

if the two objects are tied, and equals $0+1+1=n(n-1)$ if the ordering is linear. Assume the result is true for n objects and consider the case for $n+1$. Let r be number of objects ranked first at position $(r+1)/2$ (note: if $r=1$, there is a unique object in first place.) Then

$$\sum_{j=2}^{2n+2} \|P^+_{A_{n+1}}(j/2)\| = \sum_{j=2}^{r} \|P^+_{A_{n+1}}(j/2)\| + \sum_{j=r+1}^{2r+1} \|P^+_{A_{n+1}}(j/2)\| + \sum_{j=2r+2}^{2n+2} \|P^+_{A_{n+1}}(j/2)\|$$

$$= (r-1)(n+1)+(r+1)(n-r+1) + \sum_{j=2r+2}^{2n} \|P^+_{A_{n+1}}(j/2)\|.$$

The last term above corresponds to the ranking A_{n+1} after the first ranked r objects are omitted. Hence, this is a ranking of $n+1-r$ objects. Since the Blin representation is invariant in the labelling of the columns, it follows that

$$\sum_{j=2r+2}^{2n+2} \|P^+_{A_{n+1}}(j/2)\| = \sum_{j=2}^{2n-2r} \|P^+_{\bar{A}_{n-r+1}}(j/2)\|$$

where $\bar{A}_{n-r+1}$ is the $n-r+1$ object ranking obtained from A_{n+1} after deleting the first r objects.

According to the induction assumption we get

$$\sum_{j=2r+2}^{2n+2} \|P^+_{A_{n+1}}(j/2)\| = (n-r+1)(n-r).$$

Hence,

$$\sum_{j=2}^{2n+2} \|P^+_{A_{n+1}}(j/2)\| = (r-1)(n+1)+(r+1)(n-r+1)+(n-r+1)(n-r)$$

$$= n(n-1). \qquad \text{Q.E.D.}$$

REFERENCES

[1] Ronald D. Armstrong, Wade D. Cook and Lawrence M. Seiford, *Priority ranking and consensus formation: the case of ties*, Management Sci., 28, 6 (1982), pp. 638-648.

[2] J. P. Barthelemy and B. Monjardet, *The median procedure in cluster analysis and social choice theory*, Mathematical Social Sciences, 1 (1981), pp. 235-267.

[3] J. M. Blin, *A linear assignment formulation of the multiattribute decision problem*, Revue Française d'Automatique, Informatique et Recherche Operationnelle, 10 (1976), pp. 21-32.

[4] J. M. Blin and A. B. Whinston, *A note on majority rule under transitivity constraints*, Management Sci., 20 (1974), pp. 1439-1440.

[5] ———, *Discriminant functions and majority voting*, Management Sci., 21 (1975), pp. 557-566.

[6] Wade D. Cook and Moshe Kress, *Relationships between l^1 metrics on linear ranking spaces*, SIAM J. Appl. Math., 44 (1984), pp. 209-220.

[7] WADE D. COOK AND LAWRENCE M. SEIFORD, *Priority ranking and consensus formation*, Management Sci., 24 (1978), pp. 1721-1732.

[8] J. S. DE CANI, *Maximum likelihood paired comparison ranking by linear programming*, Biometrika, 56 (1969), pp. 537-545.

[9] L. HUBER AND J. SCHULTZ, *Maximum likelihood paired-comparison ranking and quadratic assignment*, Biometrika, 62 (1975), pp. 655-659.

[10] J. G. KEMENY AND L. J. SNELL, *Preference ranking: an axiomatic approach*, in Mathematical Models in the Social Sciences, Ginn, New York, 1962, pp. 9-23.

[11] P. SLATER, *Inconsistencies in a schedule of paired comparisons*, Biometrika, 48 (1961), pp. 303-312.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

014

A SHORT PROOF OF THE RECTILINEAR ART GALLERY THEOREM*

ERVIN GYÖRI†

Abstract. A short proof is given for the theorem that in a rectilinear, simply connected art gallery, $[n/4]$ watchmen are sufficient, where n is the number of the corners, in order for at least one to have a view of each internal point.

Key words. polyominoes, cutting polygons

AMS(MOS) subject classification. 05B50

In this note, we give a simple proof of the theorem of Kahn, Klawe and Kleitman [1] (proved subsequently also by O'Rourke [2]) that in a rectilinear, simply connected art gallery, $[n/4]$ watchmen are sufficient, where n is the number of the vertices of the boundary polygon, for at least one to be in sight of each interior point. In geometrical formulation, we consider *rectilinear* polygons in the plane, i.e., polygons, the sides of which are parallel to the orthogonal axes. The angles of a rectilinear polygon are all of 90 or 270 degrees. The corresponding vertices are called *convex* and *concave* respectively. We will prove

THEOREM 1 [1], [2]. *If P is a rectilinear polygon of n vertices then there exists a set S of $[n/4]$ points in the interior of P such that all the interior of P can be seen from S, i.e., any point of P can be connected to an element of S by a line segment that does not intersect the boundary of P.*

Actually, we will prove the stronger

THEOREM 2. *If P is a rectilinear polygon of n vertices then P can be partitioned into at most $[n/4]$ rectilinear polygons of 4 or 6 vertices.*

This theorem implies Theorem 1 since Theorem 1 obviously holds for rectilinear polygons of 4 or 6 vertices. Note that the form of polygons of 4 or 6 vertices is unique, being rectangular or in the shape of a capital L, respectively.

Proof of Theorem 2. We prove the theorem by induction on n. We have nothing to prove if $n=4$ or $n=6$. Notice that the sides of a rectilinear polygon are alternatively horizontal and vertical so that the number of the sides and the number n of the vertices are even. Let n be at least 8.

PROPOSITION 1. *We may assume that there are no two vertices of P that can be connected by a horizontal or vertical segment without crossing the boundary of P.*

Proof of Proposition 1. If we can connect two vertices of P by a horizontal or vertical segment in the interior of P, then cutting P along this segment, we obtain polygons P_1 and P_2 of n_1 and n_2 vertices, respectively, such that $n_1+n_2=n$ and we are done by the induction hypothesis.

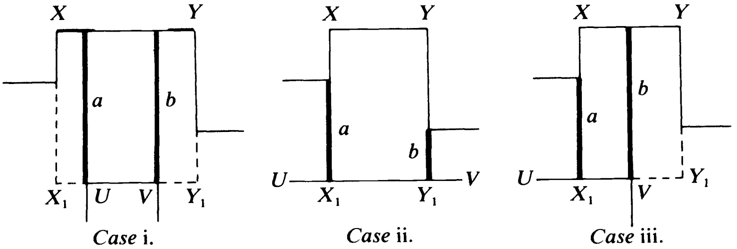
If $n=4k$ for some integer k, then we take a concave vertex V of the polygon P and cut P horizontally, say through to the boundary of P. We will then obtain polygons P_1 and P_2 of n_1 and n_2 vertices, respectively, such that $n_1+n_2=n+2$ and are done by the induction hypothesis because $n_1, n_2<n$, and

$$\left[\frac{n_1}{4}\right]+\left[\frac{n_2}{4}\right]\leqq\left[\frac{n+2}{4}\right]=\left[\frac{n}{4}\right].$$

* Received by the editors November 28, 1984, and in revised form September 1, 1985.

† Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary.

Thus, we are reduced to the case in which $n=4k+2$ for some integer k. The polygon P has $2k+3$ convex and $2k-1$ concave vertices since the sum of its interior angles is $4k\pi$. Thus, there exists two neighbouring convex vertices in P which we call X and Y. If we move the side XY orthogonally through the interior of P as far as P's boundary, let X_1 and Y_1 denote the images of X and Y, respectively, under the translation. First, suppose that upon our cutting the rectangle XYY_1X_1 out of P, we obtain two polygons. Using Proposition 1, we can easily see that we need only concern ourselves with the case in which there are one or two polygons, and if two, then X_1Y_1 intersects only one side UV of P. There are possible relations between this side UV and the segment X_1Y_1 as illustrated in Fig. 1.



Case i.

Case ii.

Case iii.

FIG. 1

Case i. X_1Y_1 contains UV.

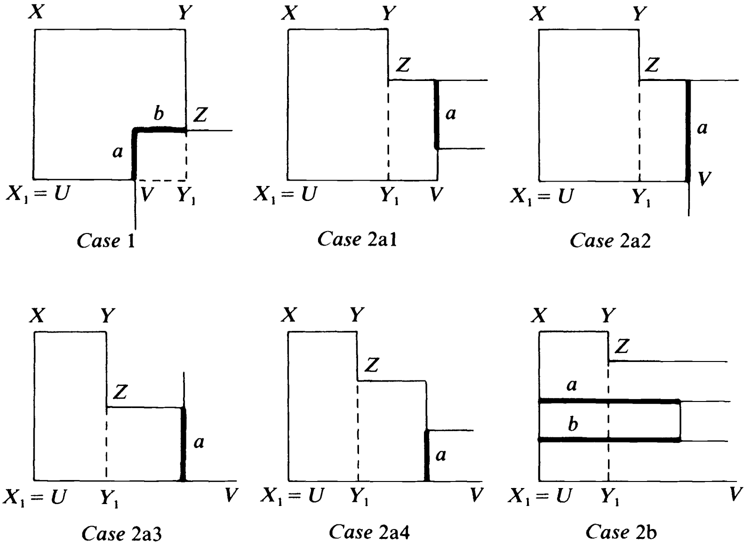
Case ii. UV contains X_1Y_1.

Case iii. X_1Y_1 and UV are overlapping, e.g. $V\in \text{int}\, X_1Y_1$, $X\in \text{int}\, UV$.

In each case, we may cut P along the thick segment a or b in Fig. 1. One or the other of these cuts yields polygons P_1 and P_2 of n_1 and n_2 vertices, respectively, with $n_1=4k_1+2$, $n_2=4k_2+2$, $k_1+k_2=k$ and we are done by the induction hypothesis.

If omission of the rectangle XYY_1X_1 leaves only one polygon, then we may assume that XX_1 is a side of P and $X_1=U$. We distinguish two cases as illustrated in Fig. 2.

Case 1. X_1Y_1 contains the side X_1V and YY_1 contains the side YZ of P.



Case 1

Case 2a1

Case 2a2

Case 2a3

Case 2a4

Case 2b

FIG. 2

Case 2. E.g. X_1Y_1 is contained in X_1V and YY_1 contains the side YZ of P. ($V \neq Y_1$ by Proposition 1.)

In Case 1, we cut P along the segments a and b obtaining polygons P_1 and P_2 of 6 and $4k-2$ vertices, respectively, in Fig. 2, and are done by the induction hypothesis.

In Case 2, we cut off an L shaped piece by cutting vertically on segment a through the first vertex. We encounter to the right of ZY, in Fig. 2, and are done by the induction hypothesis unless this cut slices P into three or more pieces. If it does so (Case 2b) one of the two possible horizontal cuts through the two interior vertices that we encountered will cut P into polygons P_1 and P_2 of $4k_1+2$ and $4k_2+2$ vertices, respectively such that $k_1+k_2=k$ and we are done by the induction hypothesis.

Acknowledgment. The author is very grateful to Jeff Kahn for suggesting the problem and for fruitful discussions, and to the referee for editorial help.

REFERENCES

[1] J. KAHN, M. KLAWE AND D. KLEITMAN, *Traditional galleries require fewer watchmen*, this Journal, 4 (1983), pp. 194-206.

[2] J. O'ROURKE, *An alternate proof of the rectilinear art gallery theorem*, J. Geometry, 21 (1983), pp. 118-130.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

015

A VARIABLE-COMPLEXITY NORM MAXIMIZATION PROBLEM*

O. L. MANGASARIAN† AND T.-H. SHIAU‡

Abstract. The decision problem associated with the problem of finding a point with largest norm in a bounded polyhedral set is shown to have a considerable range of complexity depending on the norm employed. For a p-norm with integer $p \geqq 1$, the problem is shown to be NP-complete. For the ∞-norm, the problem can be solved in polynomial time. The problem of finding an upper bound to the largest norm for any $p \in [1, \infty]$ can be solved in polynomial time by solving a single linear program.

Key words. optimization, maximum norm, complex theory, NP-complete

AMS(MOS) subject classifications. 03D15, 90C05, 90C30

1. Introduction. The problem of obtaining bounds for polyhedral sets has received considerable attention in mathematical programming [14], [15], [16], [12], [8], [9]. Part of the significance of this problem stems from the fact that the solution set to a linear program [4], [10] and to a monotone linear complementarity problem [2] is such a polyhedral set. Bounding the solution set to such problems when possible is then of practical interest. In this work we shall consider the polyhedral set X in R^n defined by

$$X := \{x \mid x \in R^n, Ax \geqq b\} \tag{1.1}$$

where A is a given $m \times n$ rational matrix and b is a given $m \times 1$ rational vector. We assume throughout this work that X is bounded. It is easy to show that a necessary and sufficient condition for X to be bounded is that

$$Y = \{y \mid y \in R^n, Ay \geqq 0, y \neq 0\} = \emptyset. \tag{1.2}$$

The problem we wish to consider here is

$$\max_{x \in X} \|x\|_p \tag{1.3}$$

where $\|\cdot\|_p$ denotes the p-norm on R^n, $1 \leqq p = \text{integer} < \infty$, defined by

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \quad \text{and} \quad \|x\|_\infty = \max_{1 \leqq i \leqq n} |x_i|.$$

We will show that while (1.3) can be solved in polynomial time for $p = \infty$, the decision problem associated with it is NP-complete [6], [11] for integer $p \geqq 1$. Since it is widely believed that *no* NP-complete problem can be solved in polynomial time (the famous conjecture $P \neq$ NP in computational complexity theory), the difference in the difficulty between $p = \infty$ and all other integer $p \geqq 1$ is enormous. (The standard complexity theory terms used here are defined in § 4.) In fact we can summarize the complexity situation for our problem (1.3) as shown in Table 1.

We note in passing that the *minimization* problem $\min_{x \in X} \|x\|_p$ is by contrast a much simpler convex programming problem for $p \in [1, \infty]$. In fact for $p = 1$ and ∞ it can be solved by standard linear programming techniques [4], [10] or by a polynomial time

* Received by the editors December 26, 1984, and in revised form September 1, 1985. This work was sponsored by the U.S. Army under contract DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under grants MCS-8200632 and DMS-8210950, Mod. 1.

† Computer Sciences Department, University of Wisconsin, Madison, Wisconsin 53706.

‡ Department of Computer Science, University of Missouri-Columbia, Columbia, Missouri 65211.

algorithm e.g. [7]. For $p=2$ the problem is a convex quadratic program which can be solved by standard techniques e.g. [2] or by a polynomial time algorithm [3].

In the following sections of this paper we will show how each of the problems of Table 1 is solved and its complexity. Section 2 deals with finding an upper bound to (1.3) for $p\in[1,\infty]$. Section 3 deals with problem (1.3) for $p=1$ and ∞ while § 4 deals with the cases of integer $p\geqq 1$.

TABLE 1
Complexity of $\max_{x\in X}\|x\|_p$ *and its method of solution.*

Problem	Complexity	Known method of solution
1. Find an upper bound to $\max_{x\in X}\|x\|_p$ for any $p\in[1,\infty]$	P (Deterministic polynomial time)	Single linear program
2. $\max_{x\in X}\|x\|_\infty$	P	$2n$ linear programs
3 $\max_{x\in X}\|x\|_1$	NP-complete (Nondeterministic polynomial time)	2^n linear programs
4. $\max_{x\in X}\|x\|_p$ Integer $p\geqq 2$	NP-complete	Vertex enumeration

2. Bounding $\max_{x\in X}\|x\|_p$. It is useful to know that for any $p\in[1,\infty]$, p not necessarily an integer, an upper bound to the solution of the nonconvex problem $\max_{x\in X}\|x\|_p$ can be obtained by solving a single linear program (Theorem 2.1 below). This is a useful result since we show (§ 4) that the problems $\max_{x\in X}\|x\|_p$ for integer $p\geqq 1$ are intractable NP-complete problems. When X is contained in the nonnegative orthant $R^n_+:=\{x\,|\,x\in R^n, x\geqq 0\}$ it is evident that a solution to the 1-norm problem $\max_{x\in X}\|x\|_1$ is easily obtained by the single linear program

$$\max_{x\in X\cap R^n_+} ex \tag{2.1}$$

where e is a vector of ones. However when $x\not\subset R^n_+$, as may be the case here, solution of $\max_{x\in X}\|x\|_1$ may take 2^n linear programs, as shown in § 3. In fact we will show in § 4 that the problem $\max_{x\in X}\|x\|_1$ is NP-complete. However, merely obtaining an upper bound to $\max_{x\in X}\|x\|_p$ for any $p\in[1,\infty]$ will take at most a single linear program as shown by the following result.

THEOREM 2.1. *Let X be nonempty and bounded, let*

$$B:=(A^TA)^{-1}A^T, \qquad d:=Bb \tag{2.2}$$

and let $B_{\cdot j}$ denote the jth column of B. Then for any $p\in[1,\infty]$ and any $x\in X$

$$\|x\|^p_p\leqq \max_{1\leqq j\leqq m}\{\|d\|^p_p, \|\gamma B_{\cdot j}+d\|^p_p\} \tag{2.3}$$

where γ is the maximum value of the following solvable linear program

$$\gamma:=\max_{x,y}\{ey\,|\,x\in R^n, y\in R^m, Ax-y=b, y\geqq 0\}. \tag{2.4}$$

Proof. Note first that the boundedness condition (1.2) implies the linear independence of the columns of A and hence the nonsingularity of A^TA. In addition the nonemptiness and boundedness of X implies the solvability of the linear program

(2.3). Hence

$$\begin{aligned}\max_{x\in X}\|x\|_p&=\max_{x,y}\{\|x\|_p\,|\,x\in R^n, y\in R^m, Ax-y=b, y\geqq 0\}\\&=\max_{x,y}\{\|x\|_p\,|\,x=By+d, (AB-I)(y+b)=0, y\geqq 0, ey\leqq\gamma\}\\&\leqq\max_{x,y}\{\|x\|_p\,|\,x=By+d, y\geqq 0, ey\leqq\gamma\}\\&=\max_{y}\{\|By+d\|_p\,|\,y\geqq 0, ey\leqq\gamma\}\\&=\max_{1\leqq j\leqq m}\{\|d\|_p, \|\gamma B_{\cdot j}+d\|_p\}.\end{aligned}$$

where the last equality follows from the fact that the maximum of a convex function on a bounded polyhedral set is attained at a vertex [13, Cor. 32.3.4]. □

Note that if a lower bound to $\max_{x\in X}\|x\|_p$ is also desired, then we have the following.

COROLLARY 2.2. *Under the assumptions of Theorem* 2.1 *we have that*

$$\|B\bar{y}+d\|_p\leqq\max_{x\in X}\|x\|_p$$

where $\bar{y}$ *is a solution of the linear program* (2.4).

Since by Khachian's result [7] a linear program is solvable in polynomial time in the size of the problem, and since the algebraic operations prescribed in (2.3) can all be performed in polynomial time, the following holds.

COROLLARY 2.3. *The bound* (2.3) *can be computed in time which is polynomial in the size of A and b.*

We note that the bound (2.3) of Theorem 2.1 may be sharp as evidenced by the following example.

Example 2.4.

$$A=\begin{pmatrix}-2 & 1\\ 5 & 1\\ 1 & -4\end{pmatrix},\qquad b=\begin{pmatrix}-10\\ -10\\ -2\end{pmatrix}.$$

For this example it is easy to verify that

$$\max_{x\in X}\|x\|_p=10\quad\text{for } p=1,2\text{ and }\infty,\quad \gamma=42,$$

$$B=\begin{pmatrix}-.0649 & .1688 & .0260\\ .0519 & .0649 & -.2208\end{pmatrix},\qquad d=\begin{pmatrix}-1.0909\\ -.7273\end{pmatrix}.$$

Computing the bound (2.3) of Theorem 2.1 gives for $p=1,2$ and ∞

$$\max_{1\leqq j\leqq 3}\{\|d\|_p, \|\gamma B_{\cdot j}+d\|_p\}=10.$$

3. $\max_{x\in X}\|x\|_p$ for $p=\infty$ and 1. It is rather obvious that the problem $\max_{x\in X}\|x\|_\infty$ can be solved by maximizing the absolute value of each component of x separately subject to x in X. This leads to the following.

PROPOSITION 3.1. *The problem* $\max_{x\in X}\|x\|_\infty$ *can be solved by solving the* $2n$ *linear programs*

$$\max_{1\leqq i\leqq n}\max\{\pm x_i\,|\,x\in R^n, Ax\geqq b\}. \tag{3.1}$$

Since each linear program can be solved in polynomial time [7] we have the following.

COROLLARY 3.2. *The problem* $\max_{x\in X}\|x\|_\infty$ *can be solved in time which is polynomial in the size of A and b.*

Since the problem $\max_{x\in X}\|x\|_1$ is equivalent to $\max_{x\in X}\sum_{i=1}^n |x_i|$, its solution can be obtained by solving 2^n linear programs as follows.

PROPOSITION 3.3. *The problem* $\max_{x\in X}\|x\|_1$ *can be solved by solving the* 2^n *linear programs*

$$\max_{v\in V}\max_{x}\{vx\,|\,x\in R^n,\ Ax\geqq b\} \tag{3.2}$$

where V is the set of 2^n *vertices of the cube in* R^n *defined by*

$$\{v\,|\,v\in R^n,\ -e\leqq v\leqq e\}, \tag{3.3}$$

where e is a vector of ones.

While $2n$ linear programs can be solved in a reasonable amount of time for intermediate-sized problems, solving 2^n linear programs is intractable even for n as small as 15. It is even worse for general $p\in(1,\infty)$ if we try to enumerate the vertices of X for finding the maximal p-norm, for the number of vertices can be as much as $\binom{m}{n}$ which, by Stirling's formula, is bounded below by an exponential in n for $m\geqq(1+\varepsilon)n$ for any constant positive ε. One may try to find other algorithms that are computationally effective. Unfortunately, as shown in the next section, problem (1.3) with $p\neq\infty$ is no easier than the partition problem (see (4.1) below) which is *inherently intractable.*

4. The intractibility of the norm maximization problem for $p\neq\infty$. We begin this section with some basic concepts of complexity theory [6], [11]. Problem A *reduces* (in polynomial time) to problem B, denoted by $A\propto B$, iff the following holds: If there is a polynomial time algorithm for B, then one can construct a polynomial time algorithm for A using the algorithm for B as a subroutine. Problems A and B are *polynomially equivalent* iff $A\propto B$ and $B\propto A$. An NP-*complete* problem is one which is polynomially equivalent to any one of the standard intractable problems such as the *satisfiability, partition,* or *travelling salesman problems* [6], [11]. These problems are considered intractable because any known algorithm which solves any one of them requires, in the worst case, an amount of time which is not bounded above by any polynomial in problem size. An NP-*hard* problem is any problem such that all problems in NP reduce to it in polynomial time. For details see [6, Chap. 5]. Thus an NP-hard problem is at least as difficult as an NP-complete problem. We will now show that our norm maximization problem (1.3) is NP-hard for $p\neq\infty$ by reducing the following NP-complete *partition problem* to it:

(4.1) Given integers $c_1, c_2, \cdots, c_n$, is there a set $S\subset\{1,2,\cdots,n\}$ such that

$$\sum_{j\in S} c_j=\sum_{j\notin S} c_j\quad ?$$

THEOREM 4.1. *The norm maximization problem* (1.3) *is* NP-*hard for* $p\in[1,\infty)$.

Proof. We will show this by reducing (4.1) to (1.3). Let $p\in[1,\infty)$. We first reduce (4.1) to the following problem:

(4.2) Given integers $c_1, c_2, \cdots, c_n$, is there an $x\in R^n$ such that:

$$\sum_{i=1}^n c_ix_i=0,\quad -1\leqq x_i\leqq 1,\quad 1\leqq i\leqq n,\quad \|x\|_p^p\geqq n\quad ?$$

It is easy to see that (4.1) has a solution S iff (4.2) has a solution x with $|x_i|=1$ for $1\leqq i\leqq n$ and $x_i=1$ for $i\in S$ and $x_i=-1$ for $i\notin S$. Now it is easy to see that (4.2) can be reduced to an instance of problem (1.3) by defining

$$A:=\begin{pmatrix} I \\ -I \\ c^T \\ -c^T \end{pmatrix},\qquad b:=\begin{pmatrix} -e \\ -e \\ 0 \\ 0 \end{pmatrix}$$

and answering the question:

$$\text{(4.3)}\qquad \text{Is}\max\{\|x\|_p^p \mid x\in R^n,\, Ax\geqq b\}\geqq n\quad ?$$

Hence if we can solve (1.3) in polynomial time we can solve each of (4.3), (4.2) and (4.1) in polynomial time. Hence (4.1)$\propto$(1.3) and (1.3) is NP-hard. □

We go on to show now that our norm maximization problem (1.3) is in fact NP-complete for integer $p\neq\infty$. In order to do this, we introduce additional concepts from complexity theory. A *nondeterministic* algorithm is an algorithm which at each step has a finite number of moves from which to choose (instead of only one for deterministic algorithms) and it solves a problem in a finite sequence of choices leading to a correct answer. NP is the class of problems solvable by a nondeterministic algorithm in polynomial time, including (4.1) and all other NP-complete problems. In fact NP-complete problems are the class of most difficult problems in NP in the sense that each problem in NP reduces in polynomial time to each NP-complete problem. By Cook's theorem [1], [6], [11], all we need to show for (1.3) to be NP-complete is that it is NP-hard (which we already have done in Theorem 4.1) and that it is in the class NP, which we proceed to do now. In order to do that, we introduce the following *decision* problem related to our optimization problem (1.3):

(4.4) Given A, b with rational entries satisfying (1.2), and nonzero integers r, s, p, is there a vector x in R^n such that

$$Ax\geqq b,\qquad \|x\|_p^p\geqq\frac{r}{s}\quad ?$$

Note that in the proof of Theorem 4.1 we have already established that the decision problem (4.4) is NP-hard, because we reduced the partition problem (4.1) to (4.2) which is an instance of (4.4). We will now first show that (4.4) is in NP and hence it is NP-complete. Then we will show that an optimization problem (1.3) is polynomially equivalent to the NP-complete decision problem (4.4). Note that condition (1.2) which is imposed on problem (4.4) which is a necessary and sufficient condition for the boundedness of X, plays an essential role in Proposition 4.2 below which establishes that (4.4) is in NP.

PROPOSITION 4.2. *Problem* (4.4) *is in* NP *for integer* $p\geqq 1$.

Proof. It follows by the convexity of the norm and the boundedness of X by (1.2) [13], that $\|x\|_p^p\geqq r/s$ for some $x\in X$ iff $\|v\|_p^p\geqq r/s$ for some vertex v of X. Moreover, v is a vertex iff there is a $J\subset\{1,2,\cdots,m\}$, $|J|=n$ such that v is the unique solution of $A_ix=b_i$, $i\in J$, and $A_jx\geqq b_j$ for $j\notin J$. Consequently we can prescribe the following nondeterministic algorithm for solving (4.4).

ALGORITHM 4.3.

```
(i) choose J, a subset of {1, 2, ···, m} with cardinality n.
(ii) Solve $A_ix=b_i$, $i\in J$ for one x, or conclude that the system is inconsistent.
(iii) if solution x found and $A_jx\geqq b_j$ for $j\notin J$ and $\|x\|_p^p\geqq r/s$ then print x; success;
      else failure; endif.
```

Step (ii) can be performed in polynomial time (e.g. by Gaussian elimination). Since we have assumed that p is an integer, $\|x\|_p^p$ can be evaluated in polynomial time. Hence Algorithm 4.3. is a polynomial time algorithm and (4.4) is in NP. □

In standard terminology, the terms NP and NP-complete refer to decision problems only but not to optimization problems. Now we show that the NP-complete decision problem (4.4) and our optimization problem (1.3) are polynomially equivalent. First it is obvious that if one can solve the optimization problem (1.3), then one can answer the decision problem (4.4). The reverse is usually done by a binary search technique showing that the optimization problem can be solved by a polynomial number of decision problems. This is all rather obvious for discrete combinatorial problems, but not for our continuous problem (1.3). To do this here, we shall use arguments similar to those of Khachian [7]. Define

$$L := \sum_{i,j=1}^{m,n} \log_2(|A_{ij}|+1) + \sum_i \log_2(|b_i|+1) + \log_2(nm+1) + \log_2(p+1).$$

L is the total length of binary digits representing the input A, b, n, m, p of problem (1.3).

THEOREM 4.4. *For any integer $p \geqq 1$, problem* (1.3) *is polynomially equivalent to the* NP-*complete decision problem* (4.4).

Proof. Since an optimal solution of (1.3) is at a vertex of X [13], such a vertex can be written by Cramer's rule as $(D_1/D, D_2/D, \cdots, D_n/D)^T$, where D and D_i are determinants of submatrices of $[A \ \ b]$. Hence

(i) For any vertex $v=(D_1/D, \cdots, D_n/D)^T$, $|D|<2^L$, $|D_i|<2^L$, $\|v\|_p^p<2^{pL}$. (See [5] for details.)

(ii) For any two distinct vertices $\|v\|_p \neq \|w\|_p$, $v=(D_1/D, \cdots, D_n/D)^T$, $w=(B_1/B, \cdots, B_n/B)^T$ it follows that

$$\left|\|v\|_p^p - \|w\|_p^p\right| = \left|\frac{|D_1|^p+\cdots+|D_n|^p}{|D|^p} - \frac{|B_1|^p+\cdots+|B_n|^p}{|B|^p}\right| \geqq \frac{1}{|D|^p|B|^p} > 2^{-2pL}.$$

Hence we can reduce (1.3) to (4.4) by binary search on the interval $[0, 2^{pL}]$ until the range is less than 2^{-2pL}. Since each iteration reduces range by half, $3pL$ iterations will do that by the following:

ALGORITHM 4.5
```
(i)   $l \leftarrow 0, u \leftarrow 2^{pL}$.
(ii)  for $i = 1$ to $3pL$ do
(iii)     solve the decision problem (4.4) for input $A$, $b$, $r/s = \frac{1}{2}(l+u)$
(iv)      if answer is yes then $l \leftarrow r/s$ else $u \leftarrow r/s$ endif
(v)   end for
```

If (iii) can be done in polynomial time, then (i) to (v) can be done in polynomial time. After (v), we know that there exists an $x \in X$ such that $l = u - 2^{-2pL}$, $\|x\|_p^p \geqq l$, whereas there is no $x \in X$ such that $\|x\|_p^p \geqq u$. Hence if we use Algorithm 4.3 with input $r/s = l$, A and b, the x printed in step (iii) of Algorithm 4.3 is an exact vertex solution of (1.3) obtained in polynomial time. Hence (1.3) reduces to (4.4). □

REFERENCES

[1] S. A. COOK, *The complexity of theorem proving procedures*, Proc. 3rd ACM Symposium on the Theory of Computing, 1971, pp. 151-158.

[2] R. W. COTTLE AND G. B. DANTZIG, *Complementary pivot theory of mathematical programming*, Linear Algebra Appl., 1 (1968), pp. 103-125.

[3] S. J. CHUNG AND K. G. MURTY, *Polynomially bounded ellipsoid algorithms for convex quadratic programming* in Nonlinear Programming 4, O. L. Mangasarian, R. R. Meyer and S. M. Robinson, eds., Academic Press, New York, 1981, 439-485.

[4] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton Univ. Press, Princeton, NJ, 1963.

[5] P. GÁCS AND L. LOVÁSZ, *Khachian's algorithm for linear programming*, Mathematical Programming Study, 14, 1981, pp. 61-68.

[6] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractibility: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, CA, 1979.

[7] L. G. KHACHIAN, *A polynomial algorithm in linear programming*, Dokl. Akad. Nauk SSR 244, 5, (1979), pp. 1093-1096; Soviet Math. Doklady, 20 (1979), pp. 191-194.

[8] O. L. MANGASARIAN, *Characterizations of bounded solutions of linear complementarity problems*, Mathematical Programming Study, 19 (1982), pp. 153-166.

[9] ———, *Simple computable bounds for solutions of linear complementarity problems and linear programs*, Mathematical Programming Study, 25, 1985, pp. 1-12.

[10] K. G. MURTY, *Linear Programming*, John Wiley, New York, 1983.

[11] C. H. PAPADIMITRIOU AND K. STEIGLITZ, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, NJ, 1982.

[12] S. M. ROBINSON, *A characterization of stability in linear programming*, Oper. Res., 25 (1977), pp. 435-447.

[13] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.

[14] A. C. WILLIAMS, *Marginal values in linear programming*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 82-94.

[15] ———, *Boundedness relations for linear constraint sets*, Linear Algebra Appl., 3 (1970), pp. 129-141.

[16] ———, *Complementarity theorems for linear programming*, SIAM Rev., 12 (1970), pp. 135-137.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

016

KELDYSH CHAINS AND FACTORIZATIONS OF MATRIX POLYNOMIALS*

EIVIND STENSHOLT†

Abstract. The paper studies factorizations $a(\lambda)=b(\lambda)\cdot c(\lambda)$ of $n\times n$ matrix polynomials, with emphasis on linear $b(\lambda)$. It is assumed that $\det a(\lambda)$ factorizes into scalar polynomials of degree 1. The tool is the concept of Keldysh chains (generalized Jordan chains). In Theorem 1 the linear independence of eigenvectors of a matrix is generalized to Keldysh chains.

It is known that there is an $nm\times nm$ matrix polynomial $Kt(m, a(\lambda))$ such that the Keldysh chains of length $\leqq m$ with respect to an "eigenvalue" ω ($\det a(\omega)=0$) may be identified with the nonzero vectors in the nullspace of $Kt(m, a(\omega))$. The main technique of the paper is to exploit the multiplicative property $Kt(m, a(\lambda))=Kt(m, b(\lambda))\cdot Kt(m, c(\lambda))$ together with the mentioned generalization. Improved factorization results are obtained when $\det a(\lambda)$ has more than $(n-1)\cdot\deg a(\lambda)$ different zeros. Also a number of known facts are derived in a simpler way than before.

Key words. matrix polynomials, Keldysh chains, Jordan chains

AMS(MOS) subject classification. 15A54

Introduction. Consider an $n\times n$ matrix polynomial

$$a(\lambda)=\sum_{i=0}^{p} a_i\lambda^{p-i},\qquad a_i\in M_n(K),\quad a_0\neq 0 \tag{0.1}$$

where $M_n(K)$ is the ring of $n\times n$ matrices over a field K and the indeterminate λ commutes with the matrices of $M_n(K)$. We are concerned with factorizations

$$\begin{aligned} a(\lambda)=b(\lambda)\cdot c(\lambda)\quad \text{where } b(\lambda)=\sum_{i=1}^{q} b_i\lambda^{q-i},\quad c(\lambda)=\sum_{i=1}^{r} c_i\lambda^{r-i}\\ \text{and } b_i, c_i\in M_n(K),\quad b_0\neq 0,\quad c_0\neq 0.\end{aligned} \tag{0.2}$$

We assume $q+r=p$, which is certainly the case if b_0 or c_0 is nonsingular. If a_0 is singular, it may happen that the scalar polynomial $\det a(\lambda)$ is constant. We deal only with cases where $\det a(\lambda)$ has positive degree.

As described in [3, p. 4], factorizations of matrix polynomials play a role in many contexts. Most familiar is perhaps the replacement of a system of differential equations, $a(D)x=u$, by two systems of lower order, $b(D)y=u$ and $c(D)x=y$. The notion of complete sets (pairs) of linear factors (Remark 5.12) is used to discuss the solutions of 2. Order systems in [3, pp. 75–79], and to develop a generalization of the Lagrange interpolation formula in [1].

Factorizations of shift operator polynomials may be used to remove (almost) nonstationarities from multivariate time series [4]. Other situations mentioned in [3, p. 4] as requiring such factorizations are decouplings in systems theory, and the study of Toeplitz matrices and Wiener-Hopf equations.

The observation that (0.2) implies

$$\det a(\lambda)=\det b(\lambda)\cdot\det c(\lambda) \tag{0.3}$$

is a natural starting point for investigations. To exploit (0.3) we assume that K is big enough to split $\det a(\lambda)$, i.e. that

$$\det a(\lambda)=\kappa\cdot\prod_{i=1}^{s}(\lambda-\omega_i)^{u_i} \tag{0.4}$$

* Received by the editors March 11, 1985.

† Norges Handelshøyskole, 5035 Bergen, Norway.

where $\kappa, \omega_i \in K$, $\omega_i \neq \omega_j$ if $i \neq j$, and $u_i > 0$. If a_0 is nonsingular, then $\kappa = \det a_0$ and $\sum u_i = np$.

1. The Keldysh chains. With $a(\lambda)$ as in (0.1) we define $a^{[k]}(\lambda)$ for all integers k by

$$a^{[k]}(\lambda) = \sum_{i=0}^{p} \binom{p-i}{k} a_i \lambda^{p-k-i} \tag{1.1}$$

where, by convention, $\binom{x}{y} = 0$ if $y \notin \{0, 1, \cdots, x\}$. $a^{[k]}(\lambda)$ is related to the formal derivative $a^{(k)}(\lambda)$ by

$$a^{(k)}(\lambda) = k! \cdot a^{[k]}(\lambda) \tag{1.2}$$

if $k \geqq 0$. For an arbitrary positive integer m, the partitioned $nm \times nm$ matrix $Kt(m, a(\lambda))$ is defined by its (i,j)-block:

$$Kt(m, a(\lambda))_{ij} = a^{[i-j]}(\lambda), \quad 1 \leqq i \leqq m, \quad 1 \leqq j \leqq m, \quad \text{i.e.,}$$

$$Kt(m, a(\lambda)) = \begin{bmatrix} a(\lambda) & 0 & \cdots & 0 \\ a^{[1]}(\lambda) & a(\lambda) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ a^{[m-1]}(\lambda) & a^{[m-2]}(\lambda) & \cdots & a(\lambda) \end{bmatrix}. \tag{1.3}$$

These matrices have appeared previously, e.g. in [3, p. 25]. Their multiplicative property seems, however, to have been largely overlooked.

LEMMA 1. *$Kt(m, b(\lambda) \cdot c(\lambda)) = Kt(m, b(\lambda)) \cdot Kt(m, c(\lambda))$.*

Proof. With $a(\lambda) = b(\lambda) \cdot c(\lambda)$, this matrix equation is equivalent to

$$a^{[k]}(\lambda) = \sum_{j=0}^{k} b^{[j]}(\lambda) \cdot c^{[k-j]}(\lambda). \tag{1.4}$$

Let first K have characteristic 0. Then (1.2) shows that (1.4) is a consequence of the product rule of formal derivation. Substituting $a_i = \sum_j b_j c_{i-j}$ in $a^{[k]}(\lambda)$ and comparing terms on each side of (1.4), one gets the (well-known) identity

$$\sum_{j=0}^{k} \binom{x}{j} \binom{y}{k-j} = \binom{x+y}{k}.$$

This identity, in turn, implies (1.4) for an arbitrary field K. □

Let $m_j(\lambda) = I_n \lambda^j$ with $I_n =$ the $n \times n$ identity matrix. For arbitrary positive integers d, e, the partitioned $nd \times ne$ matrix $V(d, e, \lambda)$ is defined by its (i,j)-block:

$$V(d, e, \lambda)_{ij} = m_{j-1}^{[i-1]}(\lambda), \quad 1 \leqq i \leqq d, \quad 1 \leqq j \leqq e, \quad \text{i.e.,}$$

$$V(d, e, \lambda) = \begin{bmatrix} I_n & \lambda & \lambda^2 & \cdots & \lambda^{e-1} \\ 0 & I_n & 2\lambda & \cdots & (e-1)\lambda^{e-2} \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & \binom{2}{d-1}\lambda^{3-d} & \cdots & \binom{e-1}{d-1}\lambda^{e-d} \end{bmatrix}. \tag{1.5}$$

The $nd \times n$ matrix $R(d, e, \lambda)$ is defined as the right-hand block column of $V(d, e+1, \lambda)$, i.e.

$$V(d, e+1, \lambda) = [V(d, e, \lambda) | R(d, e, \lambda)]. \tag{1.6}$$

Consider a finite sequence $\delta_1, \cdots, \delta_e$ of vectors in K^n, and for convenience write $\delta_i = 0$ if $i \leqq 0$. Let $m \geqq l$. With the sequence we associate a partitioned $l \times nm$ matrix F defined by its (i, j)-block:

$$F_{i,j} = \delta_{i-j+1}, \quad 1 \leqq j \leqq l, \quad 1 \leqq j \leqq m, \quad \text{i.e.,}$$

$$F = \begin{bmatrix} \delta_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \delta_2 & \delta_1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \delta_l & \delta_{l-1} & \cdots & \delta_1 & 0 & \cdots & 0 \end{bmatrix}. \tag{1.7}$$

Lemma 2. *With $\omega \in K$, the following four statements are pairwise equivalent*:

$$\sum_{k=0}^{i-1} \delta_{i-k} a^{[k]}(\omega) = 0 \quad \textit{for } 1 \leqq i \leqq l, \tag{1.8}$$

$$(\delta_l, \delta_{l-1}, \cdots, \delta_1, 0, \cdots, 0) \cdot Kt(m, a(\omega)) = 0, \tag{1.9}$$

$$F \cdot Kt(m, a(\omega)) = 0, \tag{1.10}$$

$$[F \cdot V(m, p, \omega)] \begin{bmatrix} a_p \\ \vdots \\ a_1 \end{bmatrix} = -F \cdot R(m, p, \omega) \cdot a_0. \tag{1.11}$$

Proof. The equivalences (1.8) ⇔ (1.9) ⇔ (1.10) are immediate by the definitions (1.3) and (1.7). Using the definitions and rearranging terms, we now rewrite (1.8) successively as follows:

$$\sum_{k=0}^{i-1} \delta_{i-k} \sum_{t=0}^{p} \binom{p-t}{k} a_t \omega^{p-k-t} = \sum_{t=0}^{p} \left[\sum_{k=0}^{i-1} \delta_{i-k} \binom{p-t}{k} \omega^{p-k-t} \right] a_t = 0,$$

$$\sum_{t=1}^{p} \left[\sum_{k=0}^{i-1} \delta_{i-k} \binom{p-t}{k} \omega^{p-k-t} \right] a_t = -\sum_{k=0}^{i-1} \delta_{i-k} \binom{p}{k} \omega^{p-k} a_0,$$

$$\sum_{u=1}^{p} \left[\sum_{k=0}^{i-1} \delta_{i-k} \binom{u-1}{k} \omega^{u-k-1} \right] a_{p+1-u} = -\sum_{k=0}^{i-1} \delta_{i-k} \binom{p}{k} \omega^{p-k} a_0,$$

$$\sum_{u=1}^{p} \left[\sum_{k=0}^{i-1} F_{i,k+1} \cdot V(m, p, \omega)_{k+1,u} \right] a_{p+1-u} = -\sum_{k=0}^{i-1} F_{i,k+1} \cdot V(m, p+1, \omega)_{k+1,p+1} \cdot a_0.$$

The last l equations $(1 \leqq i \leqq l)$ may be written in matrix form as (1.11). □

Definition 1.12. The pair $(\{\delta_1, \cdots, \delta_l\}, \omega)$ will be called a Keldysh chain for $a(\lambda)$ if $\delta_1 \neq 0$ and (1.8) holds.

Remark 1.13. The matrix formulations (1.9) and (1.10) are theoretically useful to "test" whether a pair is a Keldysh chain by means of the "Keldysh test matrix" $Kt(m, a(\lambda))$. The message of (1.11) is that a knowledge of Keldysh chains for $a(\lambda)$ means a knowledge of linear equations which the entries of the coefficient matrices $a_0, \cdots, a_p$ must satisfy. Since $a(\lambda)$ and $a(\lambda) \cdot z$ have the same Keldysh chains if $z \in M_n(K)$ is nonsingular, a matrix polynomial is at most determined up to a nonsingular right hand factor by its Keldysh chains.

Remark 1.14. If $a(\lambda) = I_n \lambda - A$, $A \in M_n(K)$, then (1.8) is equivalent to $\delta_i A = \omega \delta_i + \delta_{i-1}$, $1 \leqq i \leqq l$. This means that $\delta_1, \cdots, \delta_l$ form a Jordan chain for A with respect to the eigenvalue ω (provided $\delta_1 \neq 0$) [3, p. 24, 49].

2. Keldysh chains and factorizations. Consider a monic matrix polynomial $d(\lambda)=\sum d_i\lambda^{q-i}$ of degree q (i.e. $d_i \in M_n(K)$ with $d_0=I_n$). With $a(\lambda)$ as in (0.1), we may write

$$a(\lambda)=d(\lambda)\cdot e(\lambda)+f(\lambda) \tag{2.1}$$

where $f(\lambda)$ has degree $<q$ or $f(\lambda)=0$. The division algorithm for matrix polynomials is discussed in [3, p. 89]; here, however, it is enough to check that (2.1) holds for $e(\lambda)=\sum e_i\lambda^{p-q-i}$ where the $e_i \in M_n(K)$ satisfy the recursion

$$e_i=a_i-\sum_{j=0}^{i-1} d_{i-j}e_j, \qquad 0\leqq i\leqq p-q.$$

The next two results are consequences of Lemma 1:

LEMMA 3. *If* (0.2) *holds, every Keldysh chain for* $b(\lambda)$ *is also a Keldysh chain for* $a(\lambda)$.

LEMMA 4. *Assume that*

(i) $d(\lambda)$ *is uniquely determined among the monic* $n\times n$ *matrix polynomials of degree* q *by its Keldysh chains* (*through* (1.11)), *and*

(ii) *all Keldysh chains for* $d(\lambda)$ *are also Keldysh chains for* $a(\lambda)$.

Then $a(\lambda)=d(\lambda)\cdot e(\lambda)$.

Proof. Lemma 3 is immediate by Lemma 1. To see the partial converse result in Lemma 4, we first rewrite (2.1) as

$$\begin{aligned}&a(\lambda)=d(\lambda)g(\lambda)+h(\lambda) \quad \text{where } g(\lambda)=e(\lambda)-I_n,\\ &h(\lambda)=d(\lambda)+f(\lambda).\end{aligned}$$

By Lemma 1, (1.1) and (1.3), we have

$$Kt(m,a(\lambda))=Kt(m,d(\lambda))\cdot Kt(m,g(\lambda))+Kt(m,h(\lambda)). \tag{2.2}$$

Now, if $(\delta_l,\cdots,\delta_1,0,\cdots,0)Kt(m,d(\omega))=0$, it follows by (ii) and Lemma 2 that $(\delta_l,\cdots,\delta_1,\ 0,\cdots,0)Kt(m,a(\omega))=0$. Therefore, by (2.2), $(\delta_l,\cdots,\delta_1,\ 0,\cdots,0)\cdot Kt(m,h(\omega))=0$. Because of (i), $d(\lambda)=h(\lambda)$, and $f(\lambda)=0$. □

Remark 2.3. Condition (i) is always satisfied, see e.g. [3, p. 58, Thm. 2.4] for the complex case ($K=C$). According to [3, p. 203, Cor. 7.11] the condition that $d(\lambda)$ is monic may be removed, if the notion of Keldysh chains is properly generalized (to infinite ω).

3. Linear dependence among Keldysh chains. Let $a(\lambda)$ be as in (0.1) with $\det a(\lambda)$ of positive degree.

LEMMA 5. *Assume* $\rho_1,\cdots,\rho_t\in K$ *and* $x_1,\cdots,x_t\in K^n-\{0\}$ *satisfy the following conditions*:

(i) $x_ia(\rho_i)=0$.

(ii) *If* $\rho_g=\cdots=\rho_h$, *then* $x_g,\cdots,x_h$ *are linearly independent.*

Then, if V *is a* d-*dimensional subspace of* K^n, *there are at most* dp *values of* i *such that* $x_i\in V$.

Proof. We argue by contradiction, and assume that $x_1,\cdots,x_d$ are linearly independent, while

$$x_i=\sum_{j=1}^{d} t_{ij}x_j \quad \text{with } t_{ij}\in K \quad \text{and } 1\leqq i\leqq t=dp+1. \tag{3.1}$$

Letting $e(\lambda)$ be the $d\times n$ matrix defined by

$$(3.2)\qquad e(\lambda)=\begin{bmatrix} x_1 \\ \vdots \\ x_d\end{bmatrix}\cdot a(\lambda),$$

Assumption (i) may be written (using (3.1)) as

$$(3.3)\qquad (t_{i1},\cdots,t_{id})\cdot e(\rho_i)=0,\qquad 1\leqq i\leqq dp+1.$$

If $\rho_g=\cdots=\rho_h$ for r indices $g,\cdots,h$, the vectors

$$(t_{g1},\cdots,t_{gd}),\cdots,(t_{h1},\cdots,t_{hd})$$

are linearly independent in K^d because of assumption (ii). Hence, by (3.3)

$$(3.4)\qquad e(\rho_g)=\cdots=e(\rho_h)\text{ has rank at most } d-r.$$

Since $\det a(\lambda)\neq 0$, $e(\lambda)$ has rank d (over the extended field $K(\lambda)$). We may therefore pick a nonzero $d\times d$ minor determinant $\varphi(\lambda)$ of $e(\lambda)$. $\varphi(\lambda)$ is then a polynomial in λ of degree at most dp.

Writing $\varphi_i(\lambda)$, $1\leqq i\leqq d$ for column vector i in $\varphi(\lambda)$, we have

$$\varphi(\lambda)=\det[\varphi_1(\lambda),\cdots,\varphi_d(\lambda)].$$

The product rule of formal derivation implies

$$\varphi'(\lambda)=\det[\varphi_1'(\lambda),\cdots,\varphi_d(\lambda)]+\cdots+\det[\varphi_1(\lambda),\cdots,\varphi_d'(\lambda)].$$

If $0\leqq u\leqq r-1$, $\varphi^{(u)}(\lambda)$ is therefore a sum of determinants each of which has at least $d-r+1$ columns in common with $e(\lambda)$. Hence, by (3.4)

$$\varphi^{(u)}(\rho_g)=0\quad\text{for } 1\leqq u\leqq r-1.$$

This means that $\varphi(\lambda)$ is divisible by $(\lambda-\rho_g)^r$. Consequently $\varphi(\lambda)$ is divisible by $\Pi(\lambda-\rho_i)$, a polynomial of degree $dp+1$, which is absurd. □

Let $(\{\delta_1^{(i)},\cdots,\delta_{l_i}^{(i)}\},\rho_i)$ be a Keldysh chain for $a(\lambda)$ and consider the $\sum l_i$ vectors

$$\gamma_{ij}=(\delta_j^{(i)},\cdots,\delta_1^{(i)},\cdots,0,\cdots,0)\in K^{mn},\qquad 1\leqq i\leqq t,\quad 1\leqq j\leqq l_i$$

where $m\geqq l_i$ for all i.

THEOREM 1. *Assume that if $\rho_g=\cdots=\rho_h$, then $\delta_1^{(g)},\cdots,\delta_1^{(h)}$ are linearly independent. If V is a d-dimensional subspace of K^{mn}, there are at most dp values of (i,j) such that $\gamma_{ij}\in V$.*

Proof. This follows by replacing x_i by γ_{ij} and $a(\lambda)$ by $Kt(m,a(\lambda))$ in Lemma 5. The assumption of the theorem clearly implies that the vectors $\gamma_{gj},\cdots,\gamma_{hk}$ are linearly independent $(1\leqq j\leqq l_g,\cdots,1\leqq k\leqq l_h)$. □

4. Linear $b(\lambda)$. An obvious plan for attacking the general factorization problem (0.2) is to exploit Lemmas 3 and 4. This requires some knowledge of the Keldysh chains. For a monic matrix polynomial the Keldysh chain structure is closely related to the Jordan form of the companion matrix [3, p. 4, p. 40].

A purpose of the present paper, however, is to get factorization results without any elaborate techniques. This is possible in the case of linear $b(\lambda)=\lambda-B$ in (0.2), i.e.

$$(4.1)\qquad a(\lambda)=(\lambda-B)\cdot c(\lambda).$$

Comparison of coefficients shows that $a(\lambda)$ and B determine $c(\lambda)$ uniquely.

THEOREM 2. *If* $s=(n-1)p+u$ *with* $1 \leqq u \leqq p$ (s *as in* (0.4)), *the pair* $(B, c(\lambda))$ *may be chosen so that* (4.1) *holds and*

(i) B *has* n *different eigenvalues*;

(ii) *the equation* $\det c(\lambda)=0$ *has at least* $(n-1)(p-1)+(u-1)$ *different solutions. The choice may be done in at least* $(n!)^{-1}\prod_{i=0}^{n-1}(u+ip)$ *different ways.*

Proof. Let $(\{x_i\}, \omega_i)$, $1 \leqq i \leqq s$, be (length 1) Keldysh chains for $a(\lambda)$, i.e. $x_i a(\omega_i)=0$ and $x_i \neq 0$. According to Lemma 5, a linearly independent ordered subset $\{x_{i(1)}, \cdots, x_{i(n)}\}$ of $\{x_1, \cdots, x_s\}$ may be chosen in at least

$$[(n-1)p+u] \cdot [(n-2)p+u] \cdot \cdots \cdot [0 \cdot p+u]$$

different ways, since after the choice of $x_{i(1)}, \cdots, x_{i(j)}$, at most jp of the vectors $x_1, \cdots, x_s$ have become ineligible.

Let B be the unique $n \times n$ matrix with (length 1) Jordan chains $(\{x_{i(j)}\}, \omega_{i(j)})$, $1 \leqq j \leqq n$. By Remark 1.14 this means that $\lambda - B$ has (length 1) Keldysh chains $(\{x_{i(j)}\}, \omega_{i(j)})$. According to Lemma 4, (4.1) holds for suitable $c(\lambda)$. Since $s-n=(n-1)(p-1)+(u-1)$, statement (ii) also holds. □

5. The case $s=np$. We now consider the case $u=p$ in Theorem 2. Then $s=np$ and (0.4) has np distinct zeros. This forces $\det a_0 \neq 0$. For each $j \in \{1, 2, \cdots, np\}$ choose $x_j \in K^n$ such that $x_j \neq 0$ and $x_j a(\omega_j)=0$. The 1-spaces $\langle x_j \rangle$ are uniquely determined, because if $y \cdot a(\omega_j)=0$ and $y \in K^n - \langle x_j \rangle$, the set $\{x_1, \cdots, x_{np}, y\}$ would violate Lemma 5 with $V=K^n$, $t=np+1$.

A more general case, where the companion matrix of $a(\lambda)$ has diagonal Jordan form, is discussed in [3, p. 113, Thm. 3.21]. The following collection of results is essentially a corollary of Theorem 2 and the preceding lemmas.

5.1 *There exist factorizations* $a(\lambda)=(\lambda-B) \cdot c(\lambda)$.

5.2. *To each such factorization there exists a unique subset* $S \leqq \{1, 2, \cdots, np\}$ *with* $|S|=n$ *and* $x_j B=\omega_i x_j$ *for* $j \in S$.

5.3. *The set* S *determines* B *uniquely.*

5.4. *A given subset* $S \leqq \{1, 2, \cdots, np\}$ *with* $|S|=n$ *determines a factor* $\lambda - B$ *if and only if the vectors* x_j, $j \in S$, *are linearly independent.*

5.5. *If* $T \subseteq \{1, 2, \cdots, np\}$ *and* $|T| \geqq dp+1$, *then the dimension of the span* $\langle x_j | j \in T \rangle$ *is at least* $d+1$.

5.6. *There exists a partitioning* $\{1, 2, \cdots, np\}=S_1 \cup S_2 \cup \cdots \cup S_p$ *where* $|S_i|=n$ *and the* n *vectors* x_j *with* $j \in S_i$ *are linearly independent* $(1 \leqq i \leqq p)$.

5.7. *If* N *is the number of factorizations* (5.1), *then* $p^n \leqq N \leqq \binom{np}{n}$. *The lower bound is attained if and only if the vectors* x_j, $1 \leqq j \leqq np$, *belong to* n *linearly independent spaces of dimension* 1. *The upper bound is attained if and only if the Haar condition is satisfied, i.e. for every subset* $S \leqq \{1, 2, \cdots, np\}$ *with* $|S|=n$, *the vectors* x_j *with* $j \in S$ *are linearly independent.*

5.8. *If* K *is infinite and* $i \to x_i$ *is a map* $\{1, 2, \cdots, np\} \to K^n$ *such that statement* 5.5 *holds, then there exists a monic* $n \times n$ *matrix polynomial* $a(\lambda)$ *of degree* p *and distinct* $\omega_1, \cdots, \omega_{np} \in K$ *such that* $x_i \cdot a(\omega_i)=0$, $1 \leqq i \leqq np$.

5.9. *If* K *is infinite, both bounds in* (5.7) *are attained.*

Proof. Results 5.1-5.5 are immediate by Theorem 2 and Lemmas 2, 3, 4 and 5. Statements 5.5 and 5.6 are equivalent statements because of the following theorem due to Jack Edmonds [2], valid in the more general context of matroids.

THEOREM (J. Edmonds). *Consider a triple* (φ, X, V) *where* X *is a set,* V *a vector space of finite dimension* (*matroid*) *and* $\varphi: X \to V$. *Let* p *be a natural number and* $r(U)$ *the dimension* (*rank*) *of a subspace* U *of* V. *The following two conditions on* (φ, X, V)

are equivalent:

(i) *For each subspace* U *of* V, $|\varphi^{-1}(U)| \leqq p \cdot r(U)$.

(ii) *There exists a partitioning* $X = X_1 \cup X_2 \cup \cdots \cup X_p$ *such that* $|\varphi(X_i)| = |X_i|$ *and* $\varphi(X_i)$ *is linearly independent*, $1 \leqq i \leqq p$.

The statements on upper bounds in 5.7 follow from 5.2, 5.3, 5.4; the statements on lower bounds follow from the counting argument in the proof of Theorem 2.

To prove 5.8, we have (by Lemma 2) to produce $a(\lambda)$ and $\omega_1, \cdots, \omega_{np}$ such that

$$\begin{gathered} x_i[I_n\omega_i \cdots \omega_i^{p-1}] \cdot \begin{bmatrix} a_p \\ \vdots \\ a_1 \end{bmatrix} = -x_i\omega_i^p, \quad \text{i.e.,} \\ [x_i, \omega_i x_i \cdots \omega_i^{p-1}x_i] \begin{bmatrix} a_p \\ \vdots \\ a_1 \end{bmatrix} = -\omega_i^p x_i, \qquad 1 \leqq i \leqq np. \end{gathered} \tag{5.10}$$

Since we already know the equivalence 5.5 $\Leftrightarrow$ 5.6, we may enumerate the x_i's and ω_i's in such a way that the matrices

$$G_j = \begin{bmatrix} x_{(j-1)n+1} \\ \vdots \\ x_{jn} \end{bmatrix}, \qquad 1 \leqq j \leqq p,$$

are nonsingular. Letting $D_j = \operatorname{diag}\{\omega_{(j-1)n+1}, \cdots, \omega_{jn}\}$, the requirement (5.10) may be written as

$$[G_j, D_jG_j, \cdots, D_j^{p-1}G_j] \begin{bmatrix} a_p \\ \vdots \\ a_1 \end{bmatrix} = -D_j^pG_j, \quad 1 \leqq j \leqq p, \quad \text{i.e.,}$$

$$\begin{bmatrix} I_n & G_1^{-1}D_1G_1 & \cdots & G_1^{-1}D_1^{p-1}G_1 \\ \vdots & \vdots & & \vdots \\ I_n & G_p^{-1}D_pG_p & \cdots & G_p^{-1}D_p^{p-1}G_p \end{bmatrix} \begin{bmatrix} a_p \\ \vdots \\ a_1 \end{bmatrix} = -\begin{bmatrix} G_1^{-1}D_1^pG_1 \\ \vdots \\ G_p^{-1}D_p^pG_p \end{bmatrix}. \tag{5.11}$$

The determinant of the block Vandermonde matrix in (5.11) is a polynomial in $\omega_1, \cdots, \omega_{np}$. Considering the terms including the product

$$\omega_{(n-1)p+1}^{p-1} \cdot \cdots \cdot \omega_{np}^{p-1},$$

and using induction on p, we see that the polynomial is not constant. By the infinity of K, the ω_i's may be chosen distinct and such that the determinant is nonzero.

When K is infinite, the map in 5.8 may be defined so that the Haar condition holds. Now 5.9 follows from 5.7 and 5.8. □

Remark 5.12. To each S_i in result 5.6 there is a factorization $a(\lambda) = (\lambda - B_i) \cdot c^i(\lambda)$. In the terminology of [1], $\{B_1, \cdots, B_p\}$ is a "complete set of left solvents" for $a(\lambda)$. The Keldysh chain approach combined with Edmonds' theorem provide a different proof for the existence of complete sets.

Remark 5.13. As a consequence of result 5.7 there are M factorizations $a(\lambda) = (\lambda - C_1)(\lambda - C_2), \cdots, (\lambda - C_p)$ with $C_i \in M_n(K)$, $1 \leqq i \leqq p$, where

$$(p!)^n \leqq M \leqq (np)! \cdot (n!)^{-p}.$$

The lower bound is an improvement on the number $\prod_{j=0}^{p-1}(jn+1)$ given in [3, p. 114, Cor. 3.22].

Acknowledgments. I am grateful to the referee for valuable remarks to an earlier version, and to Helge Tverberg for the reference [2].

REFERENCES

[1] J. E. DENNIS, Jr., J. F. TRAUB AND R. P. WEBER, *The algebraic theory of matrix polynomials*, SIAM J. Numer. Anal., 13 (1976), pp. 831-845.

[2] JACK EDMONDS, *Minimum partition of a matroid into independent subsets*, Nat. Bur. Stand. B Math. Math. Phys., 69B (1965), pp. 67.

[3] I. GOHBERG, P. LANCASTER AND L. RODMAN, *Matrix Polynomials*, Academic Press, New York, 1982.

[4] E. STENSHOLT AND D. TJØSTHEIM, *Factorizing multivariate time series operators*, J. Multivar. Anal., 11 (1981), pp. 244 ff.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

017

THE DISTRIBUTION OF PREFIX OVERLAP IN CONSECUTIVE DICTIONARY ENTRIES*

RODICA SIMION† AND HERBERT S. WILF‡

Abstract. We consider the family $\Delta(\mathbf{m}; \mathscr{A})$ of all dictionaries, over an alphabet $\mathscr{A}$, that have given numbers m_i of words of each length $i = 1, 2, \cdots$. We find the probability distribution of the length of the maximal common prefix of two consecutive words in dictionaries $\mathscr{D} \in \Delta$, and the asymptotic behavior of the average length of those common prefixes. In the case of dictionaries of D words, all of the same length, the size of the average prefix overlap is "near" $\log_A D$ $(A = |\mathscr{A}|)$.

Key words. prefix, overlap, dictionary, compression

1. Introduction. Consider the following dictionary $\mathscr{D}$ of 8 words:

RUMMAGE
RUMOR
RUMPLE
RUNABOUT
RUSSET
RUST
RUSTIC
RYE

The *overlap* of two consecutive words in $\mathscr{D}$ is the length of their maximal common prefix. The overlap $\Omega(\mathscr{D})$ of the dictionary $\mathscr{D}$ is the sum of the overlaps of all pairs of consecutive words in $\mathscr{D}$. In the example above we have $\Omega(\mathscr{D}) = 18$.

One of the devices that is used to compress dictionaries is the suppression of maximal common prefixes, and their replacement with some encoding of the length of the suppressed string. Thus we could compress the dictionary above to

RUMMAGE, 3OR, 3PLE, 2NABOUT, 2SSET, 3T, 4IC, 1YE

In this paper we will study the combinatorics and asymptotics of the distribution of overlaps in dictionaries of given type. We will *not* be concerned here with algorithms for retrieving the compressed information. An extensive discussion of such algorithms, together with a number of results that are closely related to ours, appears in Knuth [1], particularly § 6.

Let $\mathbf{m} = (m_1, m_2, \cdots)$ be a given sequence of nonnegative integers, only finitely many of which are nonzero. Let $\mathscr{A} = (a, b, c, \cdots)$ be an alphabet of A letters. A *dictionary of type* $\mathbf{m}$ is a lexicographically ordered list of words over $\mathscr{A}$, of which exactly m_i are of length i, for each $i = 1, 2, \cdots$. We will write $|\mathbf{m}| = \sum m_i = D$ for the number of words in such a dictionary.

Next we describe the probability space that is the setting for our study. Consider the ensemble $\Delta(\mathbf{m})$ of all possible dictionaries of type $\mathbf{m}$ over $\mathscr{A}$. Associated with each $\mathscr{D} \in \Delta(\mathbf{m})$ and each integer $l \geqq 0$ there is the probability $p(\geqq l, \mathscr{D}, \mathbf{m})$ that two consecutive words of $\mathscr{D}$ overlap in $\geqq l$ initial letters. If we average $p(\geqq l, \mathscr{D}, \mathbf{m})$ over all $\mathscr{D} \in \Delta(m)$,

* Received by the editors February 19, 1985.

† Southern Illinois University, Carbondale, Illinois 62901 and Bryn Mawr College, Bryn Mawr, Pennsylvania 19010.

‡ Department of Mathematics, University of Pennsylvania, Philadelphia, Pennsylvania 19104. The work of this author was supported in part by the U.S. Office of Naval Research.

we get $p(\geqq l, \mathbf{m})$, the probability that two consecutive words in a dictionary of type $\mathbf{m}$ overlap in at least l letters. Our first result is

THEOREM 1. *If two consecutive words are chosen randomly from a randomly chosen dictionary of type* $\mathbf{m}$, *then the probability that their maximal common prefix has length* $\geqq l$ *is*

$$(1.1)\quad p(\geqq l, \mathbf{m}) = \frac{1}{|\mathbf{m}|-1}\left\{\sum_{i\geqq l} m_i - A^l + A^l \frac{\binom{A^l-1}{m_l}\binom{A^{l+1}-A}{m_{l+1}}\cdots\binom{A^n-A^{n-l}}{m_n}}{\binom{A^l}{m_l}\binom{A^{l+1}}{m_{l+1}}\cdots\binom{A^n}{m_r}}\right\}$$

for $l = 0, 1, 2, \cdots$.

The special case in which all words have the same length occurs frequently, and is given by the following corollary of Theorem 1.

THEOREM 2. *If two consecutive words are chosen randomly from a randomly chosen dictionary* $\mathcal{D}$ *of* D *words, all of* r *letters, then the probability that their maximal common prefix has length* $\geqq l$ *is*

$$(1.2)\qquad \frac{1}{D-1}\left\{D - A^l + A^l \frac{\binom{A^r-A^{r-l}}{D}}{\binom{A^r}{D}}\right\}$$

for $l = 0, 1, 2, \cdots, r$.

In many applications it is the averages that are of the most interest. Exact formulas foı these averages follow at once from (1.1), (1.2). We study the asymptotic behavior of the average in the case where all words have the same length.

THEOREM 3. *Let* $f(r, D, A)$ *be the average size of the initial string overlap in consecutive words of dictionaries with* D *words, all of length* r. *If* $r, D \to \infty$ *in such a way that* $D = o(A^{r/2})$ *then* $f(r, D, A) \sim U_{A,D}/(D-1)$, *where*

$$(1.3)\qquad U_{m,n} = \sum_{k\geqq 2}\binom{n}{k}\frac{(-1)^k}{(m^{k-1}-1)}.$$

The asymptotic behavior of $U_{m,n}$ has been determined by de Bruijn and Knuth [1, Problem 5.2.2.50] to be

$$(1.4)\qquad U_{m,n} = n\log_m n + \frac{n(\gamma-1)}{\log m} - \frac{n}{2} + nf_{-1}(n) + O(1)$$

where f_{-1} is a small, bounded function, and γ is Euler's constant.

Therefore, with the precise meaning that we have just described, we may say roughly that *the average overlap tends to be near* $\log_A D$ *in a long dictionary of long words.*

Aside from the obvious applications to the compression of dictionaries in computer memories, these results apply to combinatorial "dictionaries" also, i.e. to lexicographically ordered lists of combinatorial structures such as sets, subsets, permutations, partitions, etc. An interpretation of the overlap, in these cases, is that it measures the complexity per step of the lexicographic sequencing algorithms that list the structures in question. We will present a number of special cases below.

2. The generating function for overlaps. In this section we will derive the generating function for the number of dictionaries of given type that have given overlap properties.

Precisely, we will prove

THEOREM 4. *The number of dictionaries of given type* $\mathbf{m}$ *over an alphabet of* A *letters in which exactly* r *pairs of consecutive words overlap on a prefix of at least* l *letters in which exactly* r *pairs of consecutive words overlap on a prefix of at least* l *letters is the coefficient of* $x_1^{m_1}x_2^{m_2}\cdots x_n^{m_n}t^r$ *in the expansion of*

$$G_l(\mathbf{x},t)=\prod_{i=1}^{l-1}(1+x_i)^{A^i}\left\{1+\frac{1}{t}\left(\prod_{j=0}^{n-l}(1+x_{l+j}t)^{A^j}-1\right)\right\}^{A^l}. \tag{2.1}$$

Proof. For any word w, let $|w|$ denote the number of letters in w. Then, to every word w in a dictionary $\mathscr{D}$ we assign a weight, $\alpha(w)$, as follows: if $|w|<l$ or if w and its successor have a common prefix of length less than l, then $\alpha(w)=x_{|w|}$, else $\alpha(w)=x_{|w|}t$. Further, the weight of $\mathscr{D}$ is $\alpha(\mathscr{D})=\prod_{w\in\mathscr{D}}\alpha(w)$. Note therefore that $\mathscr{D}$ has weight $x_1^{m_1}x_2^{m_2}\cdots x_n^{m_n}t^r$ iff $\mathscr{D}$ is of type $\mathbf{m}$, and $\mathscr{D}$ contains exactly r pairs of consecutive words whose common prefix has length $\geqq l$.

Now partition the dictionary $\mathscr{D}$ as

$$\mathscr{D}=\mathscr{D}^-\cup\mathscr{D}_{v_1}\cup\mathscr{D}_{v_2}\cup\cdots\cup\mathscr{D}_{v_{A^l}}$$

where $\mathscr{D}^-=\{w\in\mathscr{D}\,|\,|w|<l\}$ and $\mathscr{D}_v$ is the subset of words of $\mathscr{D}$ whose prefix is v, and finally, the words v_i $(i=1,\cdots,A^l)$ are the l-letter words over the alphabet $\mathscr{A}$. Note that the words in each $\mathscr{D}_{v_i}$ are lexicographically consecutive in $\mathscr{D}$ and that

$$\alpha(\mathscr{D})=\alpha(\mathscr{D}^-)\alpha(\mathscr{D}_{v_1})\cdots\alpha(\mathscr{D}_{v_{A^l}}).$$

Moreover,

$$\sum_{\mathscr{D}^-}\alpha(\mathscr{D}^-)=(1+x_1)^A(1+x_2)^{A^2}\cdots(1+x_{l-1})^{A^{l-1}}.$$

If $|D_{v_i}|=d_i$, then the exponent of t in $\alpha(\mathscr{D}_{v_i})$ is $d_i-1+\delta_{0,d_i}$. Thus,

$$\sum_{\mathscr{D}_{v_i}}\alpha(\mathscr{D}_{v_i})=1+\frac{1}{t}[(1+x_lt)(1+x_{l+1}t)^A(1+x_{l+2}t)^{A^2}\cdots-1],$$

since, for fixed v_i, a word in $\mathscr{D}_{v_i}$ is determined by the suffix that is to be appended to v_i. There are A^j possible suffixes of length j $(j=0,1,2,\cdots)$, each of which produces a word of length $l+j$.

Since, in forming a dictionary $\mathscr{D}$, $\mathscr{D}^-$ and the $\mathscr{D}_{v_i}$'s are selected independently, the generating functions can be multiplied, and the result (2.1) follows. □

From Theorem 4, Theorems 1 and 2 follow at once by identifying the coefficients of the monomials and dividing by the number of dictionaries of type $\mathbf{m}$. □

From Theorems 1 and 2 we can obtain formulas for the average overlaps by summation. We quote here the result in the case where the words are all of length r.

THEOREM 5. *Let* $f(r,D,A)$ *denote the average length of the prefix overlap in two consecutive words of dictionaries of* D *words, all of length* r. *Then*

$$f(r,D,A)=\frac{rD}{D-1}-\frac{A(A^r-1)}{(A-1)(D-1)}+\frac{1}{D-1}\sum_{l=1}^{r}A^l\frac{\binom{A^r-A^{r-l}}{D}}{\binom{A^r}{D}}. \tag{2.2}$$

3. Asymptotics of average overlap. In this section we will study the order of magnitude of the average prefix overlap in the case where the words in the dictionary all have the same length.

We need first a few properties of the function

$$\phi_D(x)=(1-x)^D-1+Dx. \tag{3.1}$$

Specifically we require the inequalities

$$0<\phi'_D(x)<D(D-1)x \qquad (0<x<1) \tag{3.2}$$

and

$$0<\phi_D(x)<\binom{D}{2}x^2 \qquad (0<x<1). \tag{3.3}$$

To prove these, first write $\psi(x)=\phi'_D(x)-D(D-1)x$. Then

$$\psi'=D(D-1)\{(1-x)^{D-2}-1\}<0.$$

Hence ψ decreases, $\psi(0)=0$, whence $\psi(x)<0$, proving the right member of (3.2). Since

$$\phi'_D=D\{1-(1-x)^{D-1}\}>0$$

the inequality (3.2) is established, and (3.3) follows by integration.

Next let $f(r, D, A)$ be the average size of the prefix overlap in dictionaries of D r-letter words over an alphabet of A letters. We will prove that

$$f(r, D, A)\leqq\frac{1}{D-1}U_{A,D} \tag{3.4}$$

where U is given by (1.3).

To do this, we have from the exact formula (2.2),

$$f(r, D, A)=\frac{rD}{D-1}-\frac{A(A^r-1)}{(D-1)(A-1)}+\frac{1}{(D-1)}\sum_{l=1}^{r}A^l\prod_{j=0}^{D-1}\left(1-\frac{A^{-l}}{1-jA^{-r}}\right). \tag{3.5}$$

Now since

$$\prod_{j=0}^{D-1}\left(1-\frac{A^{-l}}{1-jA^{-r}}\right)\leqq(1-A^{-l})^D$$

we have

$$\begin{aligned}f(r, D, A)&\leqq\frac{rD}{D-1}-\frac{A(A^r-1)}{(D-1)(A-1)}+\frac{1}{D-1}\sum_{l=1}^{r}A^l\{\phi_D(A^{-l})+1-DA^{-l}\}\\&=\frac{1}{D-1}\sum_{l=1}^{r}A^l\phi_D(A^{-l})\\&\leqq\frac{1}{D-1}\sum_{l=1}^{\infty}A^l\phi_D(A^{-l})\\&=\frac{1}{D-1}\sum_{l=1}^{\infty}A^l\sum_{j\geqq2}\binom{D}{j}(-1)^jA^{-lj}\\&=\frac{1}{D-1}U_{A,D}\end{aligned}$$

and (3.4) is proved. □

Next we consider lower bounds. Define

$$\delta=D/A^r. \tag{3.6}$$

This is the ratio of the number of words in the dictionary to the maximum number that might have been there. δ is small in many applications of interest.

Now for the product in (3.5) we have

$$\prod_{j=0}^{D-1}\left(1-\frac{A^{-l}}{1-jA^{-r}}\right)\geqq\left(1-\frac{A^{-l}}{1-\delta}\right)^{D}=\phi_D\left(\frac{A^{-l}}{1-\delta}\right)+1-D\frac{A^{-l}}{1-\delta}.$$

Consequently

$$\begin{aligned}f(r,D,A)&\geqq\frac{rD}{D-1}-\frac{A(A^r-1)}{(D-1)(A-1)}+\frac{1}{D-1}\sum_{l=1}^{r}A^l\left\{\phi_D\left(\frac{A^{-l}}{1-\delta}\right)+1-\frac{DA^{-l}}{1-\delta}\right\}\\&=-\frac{\delta Dr}{(D-1)(1-\delta)}+\frac{1}{D-1}\sum_{l=1}^{r}A^l\phi_D\left(\frac{A^{-l}}{1-\delta}\right).\end{aligned}\tag{3.7}$$

To estimate the sum, we write, say,

$$\begin{aligned}\sum_{l=1}^{r}A^l\phi_D\left(\frac{A^{-l}}{1-\delta}\right)&=\sum_{l=1}^{r}A^l\phi_D(A^{-l})+\sum_{l=1}^{r}A^l\left\{\phi_D\left(\frac{A^{-l}}{1-\delta}\right)-\phi_D(A^{-l})\right\}\\&=\Sigma_1+\Sigma_2\end{aligned}$$

By the mean value theorem and (3.2)-(3.3),

$$\begin{aligned}|\Sigma_2|&\leqq\sum_{l=1}^{r}A^l\left(\frac{A^{-l}\delta}{1-\delta}\right)|\phi_D'(\xi_l)|\qquad\left(A^{-l}<\xi_l<\frac{A^{-l}}{1-\delta}\right)\\&<\frac{\delta D^2}{1-\delta}\sum_{l=1}^{r}\frac{A^{-l}}{1-\delta}\\&<\frac{\delta D^2}{(1-\delta)^2}.\end{aligned}$$

As regards Σ_1, we write

$$\Sigma_1=\sum_{l=1}^{\infty}A^l\phi_D(A^{-l})-\sum_{l=r+1}^{\infty}A^l\phi_D(A^{-l})=U_{A,D}-\sum_{l>r}A^l\phi_D(A^{-l}).$$

If we substitute in (3.7), we find that

$$f(r,D,A)\geqq\frac{1}{D-1}U_{A,D}-\frac{1}{D-1}\sum_{l\geqq r+1}A^l\phi_D(A^{-l})+\varepsilon_1\tag{3.8}$$

in which

$$\begin{aligned}|\varepsilon_1|&\leqq\frac{r\delta D}{(D-1)(1-\delta)}+\frac{\delta D^2}{(D-1)(1-\delta)^2}\\&=O((r+D)\delta)\end{aligned}\tag{3.9}$$

uniformly.

Combining (3.4), (3.8), we have

$$0\leqq\frac{1}{D-1}U_{A,D}-f(r,D,A)\leqq\frac{1}{D-1}\sum_{l\geqq r+1}A^l\phi_D(A^{-l})-\varepsilon_1\tag{3.10}$$

together with the estimate (3.9) of ε_1.

The sum on the right can be further estimated. From (3.3),

$$\sum_{l\geqq r+1}A^l\phi_D(A^{-l})<\binom{D}{2}\sum_{l\geqq r+1}A^{-l}=\frac{\binom{D}{2}}{A^r(A-1)}<\frac{D^2}{A^{r+1}}\qquad(A\geqq 2).$$

Altogether this proves that

$$\left|\frac{1}{D-1}U_{A,D}-f(r,D,A)\right| = O((r+D)\delta)+O\left(\frac{D^2}{A^{r+1}}\right) \tag{3.11}$$

uniformly. Theorem 3 is now proved. □

If, finally, we use the analysis (1.4) of the size of U, we find

THEOREM 6. *We have the estimate*

$$f(r,D,A)=\log_A D+O((r+D)\delta)+O\left(\frac{D^2}{A^{r+1}}\right)+O_A(1) \tag{3.12}$$

in which the first two O's are uniform and the third depends on A.

4. Some combinatorial examples.

4. Some combinatorial examples. If we have some encoding of a family of combinatorial objects arranged lexicographically, then we can ask for the average prefix overlap of the list. Here are some examples.

In the two-line form of the permutations of $|n|$, two consecutive permutations "agree except for their last e letters," on the average.

If all 2^n subsets of $|n|$ are encoded by their membership lists in ascending order, the average overlap is $\sim n/2$ elements. If encoded as bit strings, the average overlap is $n-5/2+o(1)$.

If an ordered partition of n (= "composition of n") into positive parts is encoded by the list of its parts, then two lexicographically consecutive compositions agree on their first $n/2-3/2+o(1)$ parts, on average.

If an (unordered) integer partition is represented as its list of parts in *nondecreasing* order of size, then two consecutive "words" in the "dictionary" of partitions of n agree except for at most the largest two "letters."

Acknowledgment. Jon Bentley, of Bell Laboratories, was kind enough to obtain data from the UNIX dictionary which, together with the above theorems, convinced us of the (unsurprising) fact that a dictionary of the English language is nowhere near "average," in the sense of this paper. For those efforts, and for his enthusiasm for this project, we thank him.

REFERENCE

[1] D. E. KNUTH, *The art of computer programming, Vol.* 3: *Sorting and Searching*, Addison-Wesley, Reading, MA, 1973.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

018

THE DUAL VARIABLE METHOD FOR THE SOLUTION OF COMPRESSIBLE FLUID FLOW PROBLEMS*

J. BURKARDT†, C. HALL† AND T. PORSCHING†

Abstract. Discretizations of the Navier-Stokes equations describing a compressible flow problem can be viewed as systems defining flows on an associated network. This observation provides a means of economizing on their numerical solution.

Key words. Navier-Stokes, networks, dual variable, compressible fluid

AMS(MOS) subject classifications. 05C38, 65M10, 76D05

1. Introduction. The dual variable method [1] is a means of economizing on the cost of solving the linear or nonlinear systems that arise in certain discretizations of the Navier-Stokes equations. A matrix transformation is introduced which significantly reduces the size of the system which must be solved. For the finite difference discretization of the two-dimensional, incompressible Navier-Stokes equations studied in [1], [2] this reduction amounts to a factor of 3. A key element of the implementation of this transformation is the construction of a cycle vector basis for an associated network.

In this paper we extend the dual variable method to compressible flow problems. This again involves the use of network theory.

The system of partial differential equations in two spatial dimensions (x, y) and time t describing the compressible (barotropic) flow problem of concern is:

$$\frac{\partial \rho}{\partial t}+\nabla \cdot (\rho \mathbf{q})=0, \tag{1}$$

$$\rho\frac{\partial q}{\partial t}+\rho(\mathbf{q}\cdot\nabla)\mathbf{q}+\nabla p-\mu\left[\nabla^2\mathbf{q}+\frac{1}{3}\nabla(\nabla\cdot\mathbf{q})\right]=\mathbf{F} \tag{2}$$

where $\mathbf{q}=(u, v)$ is the velocity vector, p is pressure, μ is viscosity, $\mathbf{F}$ is a vector that includes elevation and wall friction effects, and the density ρ is determined by a state equation

$$\rho=\rho(p). \tag{3}$$

Equation (1) is referred to as the continuity equation and equation (2) as the momentum equation.

We assume that appropriate boundary conditions and an initial condition are specified so that (1)-(3) have a unique solution in a flow region Ω. Typical boundary conditions are the specification of the pressure or velocity on each segment of the boundary Ω.

In § 2, we present details of a discretization of (1)-(3). The matrix transformations involved in the dual variable method are given in § 3, and a network or physical interpretation of the dual variable transformation is given in § 4. Section 5 contains numerical results.

* Received by the editors April 8, 1985, and in revised form October 7, 1985. This work was supported by the Air Force Office of Scientific Research under grant 84-0131.

† Institute for Computational Mathematics and Applications, Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

2. The finite difference equations. There are several consistent finite difference discretizations that are available to approximate (1)–(2). We choose the following scheme based on the MAC placement of variables [3] in which a pressure is associated with the center of a control volume or mesh box and the component of velocity normal to a control volume side is associated with the center of that side (Fig. 1). Let U, V, P be the finite difference approximations to the mass velocities ρu, ρv, and pressure p respectively. A superscript m designates the mth time level.

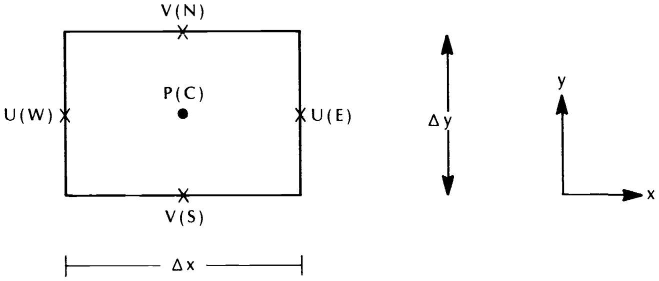


FIG. 1. *A control volume with* MAC *placement of variables and compass designations.*

In the continuity equation, we expand the time derivative via the chain rule, and use backward differencing and centered differencing on the temporal and divergence terms respectively. The discrete equation is then of the form:

$$(4)\qquad \frac{U_E^{m+1}-U_W^{m+1}}{\Delta x}+\frac{V_N^{m+1}-V_S^{m+1}}{\Delta y}+\left(\frac{\partial\rho}{\partial p}\right)_C^m\left(\frac{P_C^{m+1}-P_C^m}{\Delta t}\right)=0.$$

The momentum equations are discretized as in [1], [2] using upwind differences for the convective terms $(\mathbf{q}\cdot\nabla)\mathbf{q}$, centered differences for the pressure gradient ∇p and the viscous terms $[\nabla^2\mathbf{q}+\frac{1}{3}\nabla(\nabla\cdot\mathbf{q})]$, and backward differences for the temporal derivative $\partial\mathbf{q}/\partial t$. The finite difference system resulting from (4) and the discrete momentum equations can be written as $(N+L)$ equations

$$(5)\qquad A\mathbf{V}^{m+1}+\mathbf{W}_P^{m+1}=0,$$

$$(6)\qquad Q_m\mathbf{V}^{m+1}-\Delta t A^T\mathbf{P}^{m+1}=\mathbf{b}_1^m.$$

Here the $N\times 1$ vector $\mathbf{P}^{m+1}$, and $L\times 1$ vector $\mathbf{W}^{m+1}$ contain the unknown pressures and velocities respectively, $\mathbf{V}^{m+1}\equiv D_1\mathbf{W}^{m+1}$, D_1 a diagonal matrix with $[D_1]_{ii}=\Delta x$ if i corresponds to N or S and $[D_1]_{ii}=\Delta y$ if i corresponds to E or W (note that Δx (Δy) may vary from one column (row) of mesh boxes to the next),

$$(7)\qquad \mathbf{W}_P^{m+1}\equiv\Delta t Q_{22}^{-1}\mathbf{P}^{m+1}-\mathbf{S}^m$$

where Q_{22}^{-1} is the diagonal matrix

$$(8)\qquad Q_{22}^{-1}=\operatorname{diag}\left[\frac{\Delta x\,\Delta y}{\Delta t^2}\left(\frac{\partial\rho}{\partial p}\right)_C^m\right],$$

$$(9)\qquad [\mathbf{S}^m]_C=\frac{\Delta x\,\Delta y}{\Delta t}\left[\left(\frac{\partial\rho}{\partial p}\right)_C^m P_C^m\right]+\text{boundary mass velocities.}$$

The $N\times L$ matrix A contains 0's, 1's and -1's and can be interpreted as an incidence matrix of an associated network as described in [1], [4], [5]. The $L\times L$ matrix Q_m contains the finite difference coefficients of the discrete convective and viscous terms in the momentum equations as well as the temporal term. The $L\times 1$ vector $\mathbf{b}_1^m$ contains boundary data as well as contributions of the body force and friction terms.

If we let $\mathbf{Z}^{m+1}=[\mathbf{V}^{m+1}, \mathbf{W}_P^{m+1}]$ and

$$B=(A|I_N), \tag{10}$$

where I_N denotes the identity of order N, then the discrete continuity equation is of the generic form

$$B\mathbf{Z}^{m+1}=0. \tag{11}$$

We may think of the vector $\mathbf{W}_P^{m+1}$ as "pseudo mass flows." Equations (6) and (7) then combine to give

$$\begin{bmatrix} Q_m & 0 \\ 0 & Q_{22} \end{bmatrix}\begin{bmatrix} \mathbf{V}^{m+1} \\ \mathbf{W}_P^{m+1} \end{bmatrix}=\Delta t\begin{bmatrix} A^T \\ I_N \end{bmatrix}\mathbf{P}^{m+1}+\begin{bmatrix} \mathbf{b}_1^m \\ -Q_{22}\mathbf{S}^m \end{bmatrix}$$

or

$$\begin{bmatrix} Q_m & 0 \\ 0 & Q_{22} \end{bmatrix}\mathbf{Z}^{m+1}=\Delta t B^T\mathbf{P}^{m+1}+\mathbf{k}^m. \tag{12}$$

Thus, it is required to solve the $2N+L$ equations (11)-(12) for $\mathbf{Z}^{m+1}$ and $\mathbf{P}^{m+1}$. In the next section we show how to obtain an equivalent system from which the pressure vector $\mathbf{P}^{m+1}$ has been eliminated.

3. The dual variable transformation. The dual variable method has been used successfully in the treatment of certain finite difference and finite element discretizations of the equations of incompressible flow [1], [2]. With regard to the current system (11), (12) of compressible flow equations, the method consists of the following steps.

Step 1. Find a basis $[\mathbf{C}_1, \mathbf{C}_2, \cdots, \mathbf{C}_d]$ for the null space of B and form the $(L+N)\times d$ matrix C with $\mathbf{C}_i$ as its ith column. Then

$$BC=0 \tag{13}$$

and

$$\mathbf{Z}^{m+1}=C\mathbf{X}^{m+1} \tag{14}$$

for *some* $d\times 1$ vector $\mathbf{X}^{m+1}$.

Step 2. Substitute $\mathbf{Z}^{m+1}$ as defined by (14) into (12) to obtain

$$\begin{bmatrix} Q_m & 0 \\ 0 & Q_{22} \end{bmatrix}C\mathbf{X}^{m+1}=\Delta t B^T\mathbf{P}^{m+1}+\mathbf{k}^m. \tag{15}$$

Step 3. Multiply (15) by C^T and use the orthogonality of B^T and C^T to obtain the $d\times d$ system

$$C^T\begin{bmatrix} Q_m & 0 \\ 0 & Q_{22} \end{bmatrix}C\mathbf{X}^{m+1}=C^T\mathbf{k}^m. \tag{16}$$

The matrix transformation in (16) is called the *dual variable transformation* and (16) is called the *dual variable system.*

Step 4. Solve (16) for $\mathbf{X}^{m+1}$ and recover the velocities $\mathbf{V}^{m+1}$ and pseudo velocities $\mathbf{W}_P^{m+1}$ from (14).

Step 5. Recover the pressures from the pseudo velocities using (7), noting that Q_{22} is diagonal.

The inherent advantage of the dual variable method is the reduction in the size ($L+N$ to d) of the system to be solved at each time step. Efficient algorithms for computing well conditioned sparse bases for null spaces have been studied by Berry, Heath, Kaneko, Lawo, Plemmons and Ward [6]. This latter approach involves a matrix factorization. As we now show, C can be constructed without the need to solve any system of equations.

It is clear that the $N\times(N+L)$ matrix in (10) is of rank N. We have then that the dimension, d, of the null space of B is L, the number of columns minus the rank. Moreover, a basis for this null space is immediately provided by the columns of the matrix C defined as follows,

$$C\equiv\begin{bmatrix} I_L \\ -A \end{bmatrix}. \tag{17}$$

Substituting (17) into (16), we obtain the $L\times L$ system

$$(Q_m+A^TQ_{22}A)\mathbf{X}^{m+1}=\mathbf{b}_1^m+A^TQ_{22}\mathbf{S}^m \tag{18}$$

as the dual variable system. But by (14) and (17), $\mathbf{V}^{m+1}=\mathbf{X}^{m+1}$, and

$$\mathbf{W}_P^{m+1}=-A\mathbf{X}^{m+1}. \tag{19}$$

Hence, the unknown velocities actually satisfy (18) and the pseudo flows that are needed to recover the pressures via (7) are given by (19).

4. A network interpretation. As in the case of the discrete divergence matrix A, the augmented matrix B can also be interpreted as the incidence matrix of a directed network T. The geometric realization $G(T)$ is constructed as follows:

The *nodes* of $G(T)$ are the mesh box (control volume) centers and the *interior links* connect nodes of contiguous mesh boxes. The *boundary links* of $G(T)$ are links normal to segments of the boundary of the flow region where a pressure is specified. All links are oriented in the positive sense of the x or y axis, respectively. So far this planar network is precisely the network used in the dual variable formulation of incompressible flow problems [1], [4], [5], and A is its incidence matrix. However, we now add links which emanate from each mesh box center (node) and terminate at a fictitious node. These links are all directed toward this fictitious node. The N unknowns $[\mathbf{W}_P^{m+1}]_i$, $i=1,2,\cdots,N$ in (7) are then thought of as *pseudo-flows* on these latter links just as the L unknowns $[\mathbf{V}^{m+1}]_j$, $j=1,\cdots,L$ are flows on the links connecting mesh box centers. The $N\times(N+L)$ matrix B is the node-link incidence matrix for the network so constructed. That is,

$$[B]_{kl}=\begin{cases} +1 & \text{if link } l \text{ is directed away from node } k,\\ -1 & \text{if link } l \text{ is directed toward node } k,\\ 0 & \text{otherwise.}\end{cases}$$

Equation (11) then states that at each node the total "flow" is in balance. In Fig. 2 there are $N=14$ unknown pressures (also 14 unknown pseudo flows) and $L=22$ unknown velocities. At each of the $N=14$ nodes the sum of the flows and pseudo flows is forced to be zero.

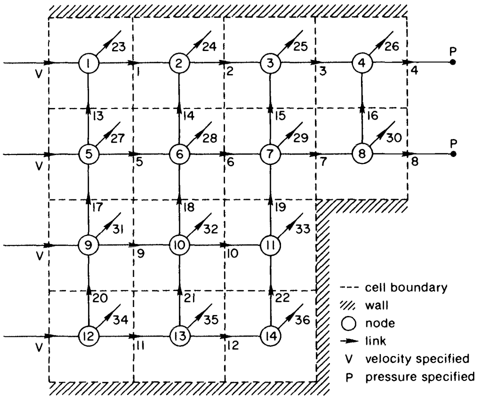


FIG. 2. *Flow region decomposed into* 14 *flow cells showing associated network of* $N=14$ *nodes and* $N+L=36$ *links.*

The 14×22 matrix A is given by

$$A=\left[\begin{array}{cccccccccccccccccccccc|c}
1&2&3&4&5&6&7&8&9&10&11&12&13&14&15&16&17&18&19&20&21&22& \\
\hline
1&0&0&0&0&0&0&0&0&0&0&0&-1&0&0&0&0&0&0&0&0&0&1\\
-1&1&0&0&0&0&0&0&0&0&0&0&0&-1&0&0&0&0&0&0&0&0&2\\
0&-1&1&0&0&0&0&0&0&0&0&0&0&0&-1&0&0&0&0&0&0&0&3\\
0&0&-1&1&0&0&0&0&0&0&0&0&0&0&0&-1&0&0&0&0&0&0&4\\
0&0&0&0&1&0&0&0&0&0&0&0&1&0&0&0&-1&0&0&0&0&0&5\\
0&0&0&0&-1&1&0&0&0&0&0&0&0&1&0&0&0&-1&0&0&0&0&6\\
0&0&0&0&0&-1&1&0&0&0&0&0&0&0&1&0&0&0&-1&0&0&0&7\\
0&0&0&0&0&0&-1&1&0&0&0&0&0&0&0&1&0&0&0&0&0&0&8\\
0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&1&0&0&-1&0&0&9\\
0&0&0&0&0&0&0&0&-1&1&0&0&0&0&0&0&0&1&0&0&-1&0&10\\
0&0&0&0&0&0&0&0&0&-1&0&0&0&0&0&0&0&0&1&0&0&-1&11\\
0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&1&0&0&12\\
0&0&0&0&0&0&0&0&0&0&-1&1&0&0&0&0&0&0&0&0&1&0&13\\
0&0&0&0&0&0&0&0&0&0&0&-1&0&0&0&0&0&0&0&0&0&1&14
\end{array}\right]$$

and the matrix $B=(A|I_{14})$ is the 14×36 incidence matrix for the network shown in Fig. 2.

As defined by (17) the matrix C may be interpreted as a *fundamental cycle basis* (cf. [7]). Each column of C is a *cycle vector* for the network T. For example, for the network of Fig. 2, the sixth column is

$$\text{(20)}\qquad [\mathrm{Col}_6(C)]_i=\begin{cases}1, & i=6 \quad \text{(link 6)},\\ -1, & i=28 \quad (L+\text{start node for link 6}),\\ 1, & i=29 \quad (L+\text{end node for link 6}),\\ 0 & \text{otherwise.}\end{cases}$$

In general, if link j is not a boundary link and is incident from node l to node k, then

$$[\mathrm{Col}_j(C)]_i=\begin{cases}1, & i=j,\, k+L,\\ -1, & i=l+L,\\ 0 & \text{otherwise.}\end{cases}$$

If link j is a boundary link, then this definition yields only two nonzero entries in the jth column of C since one of the nodes l or k does not exist.

Finally, we can conveniently use the network concept to determine the sparsity of the dual variable system (16). We observe that

$$(A^TQ_{22}A)_{jr} = (Q_{22})_{ll}A_{lr} - (Q_{22})_{kk}A_{kr} \tag{21}$$

where j, k and l are related as in Fig. 3. It follows that the right side of (21) is nonzero only when r corresponds to cycles containing those interior and boundary links that are incident to or from nodes k and l. Consequently, the jth equation of (18) is coupled only to itself and the eight other equations corresponding to the cycles containing the links shown in Fig. 4. That is, dual variable j associated with cycle j is coupled to at most 8 other dual variables associated with the cycles (or equivalently links) illustrated. Hence the coefficient matrix in (19) has at most 9 nonzero entries in each row, and, with suitable ordering of the links, may be solved as a banded matrix.

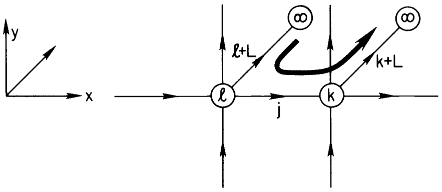


FIG. 3. *The jth cycle associated with link j. The fictitious node is labeled* ∞.

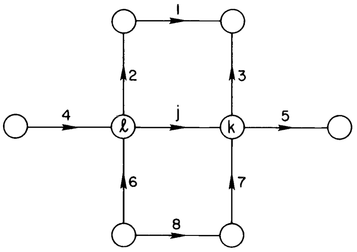


FIG. 4. *Stencil for the dual variable system* (19). *Each link corresponds to a cycle and a dual variable.*

The potential computational advantage of the dual variable method lies in the reduction of the size of the discrete Navier-Stokes system from $N+L$ to L equations and unknowns. Although the sparsity of the dual variable system (18) is less than that of the primitive system (5)-(6), the decrease is slight. The maximum number of nonzero couplings per equation increases from 7 in system (5)-(6) to 9 in system (18). Since any implementation of the method depends strongly on such imponderables as solution algorithm, data structure, computer architecture, etc., it is not possible to state unequivocally that the dual variable method will always produce dramatic reductions in running times. However, it is quite natural to expect that, "all other things being equal," the slight increase in the complexity of the dual variable system is more than offset by its decrease in size. Moreover, apart from any computational advantage the method may have, it provides additional insight into the physics of compressible flow through a novel interpretation of that phenomenon in terms of network concepts.

5. Example: aircraft cavity. Aircraft that are used for observation sometimes have cavities or compartments that open directly to the atmosphere. Figure 5 illustrates a two-dimensional model of such a cavity with dimensions as indicated. The blockage in the cavity simulates instrumentation used during the observations. The spoiler ahead

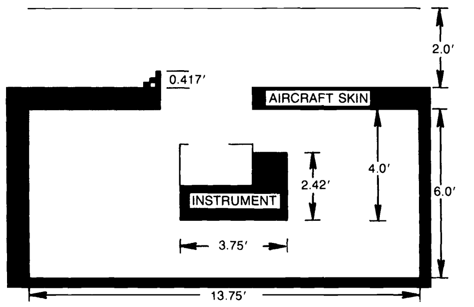


FIG. 5. *Flow region for aircraft cavity.*

of the cavity opening is used to divert the air flow so as to stabilize the flow in front of the instrumentation. We indicate a solid spoiler, although a porous spoiler may also be used. We assume the aircraft is flying at mach 0.75 and an altitude of 37,000 feet. The ambient pressure of 2.7 psi is specified at the downstream flow region boundary, and upstream the inlet velocity profile is given by:

$$u = 750\begin{cases} (y/0.417)^{1/7}, & 0 \leqq y \leqq 0.417, \\ 1, & y \geqq 0.417, \end{cases}$$

where y is the distance from the aircraft skin. All walls are assumed to be no-slip walls.

The flow region was subdivided into $N = 1167$ flow cells and there are $N = 1167$ unknown pressures. The number of unknown velocities is $L = 2206$. The primitive system (5)-(6) is of dimension $L + N = 3373$ while the dual variable system (19) is of dimension $L = 2206$. Note that the former has at most 7 nonzero elements per row, while the latter has at most 9 nonzero elements per row.

Figure 6 illustrates the streamlines in and around the cavity door. Figure 6a shows two attached vortices downstream of the spoiler or fence. These vortices separate in Fig. 6b and the downstream vortex is shed from the aircraft skin in Figs. 6c and 6d.

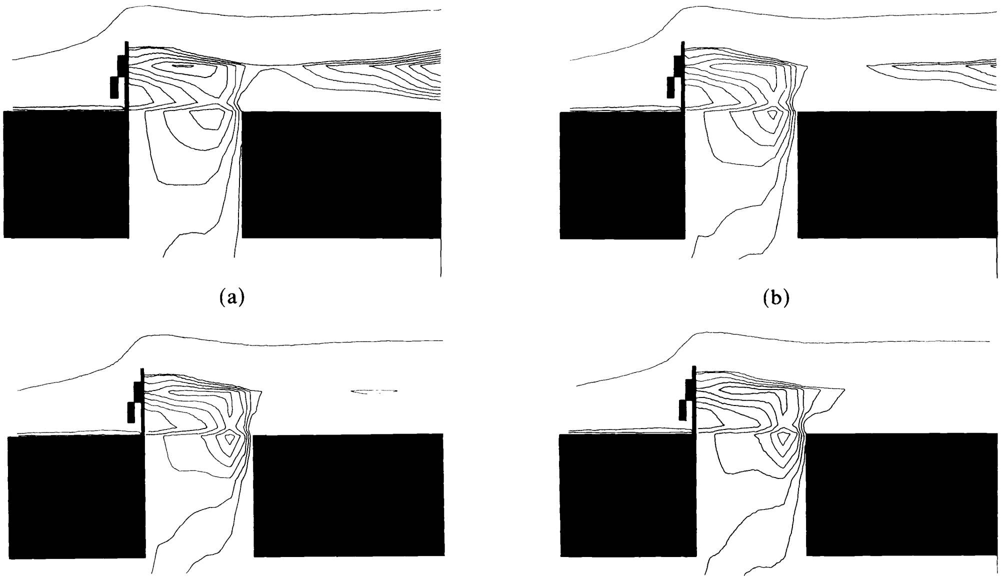


FIG. 6. (a) *Streamlines at times .005 sec*; (b) *at time .010 sec*; (c) *at time .015 sec*; (d) *at time .020 sec.*

REFERENCES

[1] R. Amit, C. A. Hall and T. A. Porsching, *An application of network theory to the solution of implicit Navier-Stokes difference equations*, J. Comp. Phys., 40 (1981), pp. 183-201.

[2] R. S. Dougall, C. A. Hall and T. A. Porsching, DUVAL: *A computer program for the numerical solution of two-dimensional, two-phase flow problems*, Volumes 1-3, Electric Power Research Institute, Report NP-2099, Palo Alto, CA, 1982.

[3] F. H. Harlow and F. W. Welch, *Numerical calculations of time dependent viscous incompressible flow of fluid with a free surface*, Phys. Fluids, 8 (1965), p. 2182.

[4] C. A. Hall, *Numerical solution of Navier-Stokes problems by the dual variable method*, this Journal, 6 (1985), pp. 220-236.

[5] T. A. Porsching, *A finite difference method for thermally expandable fluid transients*, Nucl. Sci. Eng., 64 (1977), pp. 177-186.

[6] M. Berry, M. Heath, I. Kaneko, M. Lawo, R. J. Plemmons and R. Ward, *An algorithm to compute a sparse basis of the null space*, Numerische Mathematik, to appear.

[7] N. Deo, G. M. Prabhu and M. S. Krishnamoorthy, *Algorithms for generating fundamental cycles in a graph*, ACM Trans. Math. Software, 8 (1982), pp. 26-42.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 3, July 1986

019

THE SECOND IMMANANTAL POLYNOMIAL AND THE CENTROID OF A GRAPH*

RUSSELL MERRIS†

Abstract. Let G be a graph. The Laplacian matrix, $L(G)$, is $D(G)-A(G)$, where $D(G)$ is the diagonal matrix of vertex degrees and $A(G)$ is the adjacency matrix. The second immanant of an n-by-n matrix $A=(a_{ij})$ is

$$d_2(A)=\sum_{\sigma\in S_n}\chi(\sigma)\prod_{t=1}^{n}a_{t\sigma(t)},$$

where χ is the irreducible character of S_n corresponding to the partition $(2,1,\cdots,1)$. The paper concerns various algebraic and combinatorial aspects of the coefficients $c_k(G)$ in the "d_2-polynomial"

$$d_2(xI-L(G))=\sum_{k=0}^{n}(-1)^k c_k(G)x^{n-k}.$$

For example, the effort to better understand $c_{n-1}(G)$ leads to a new definition of "centroid point". One appendix lists the d_2-polynomials for all graphs on six vertices while a second gives a BASIC program to compute the d_2-polynomial.

Key words. Laplacian matrix, permanent, centroid point, adjacency matrix, invariant

AMS(MOS) subject classifications. 05C05, 05C50, 15A15

1. Introduction. Let $G=(V,E)$ be an undirected graph without loops or multiple edges, with vertex set $V=\{1,2,\cdots,n\}$ and edge set E. The *Laplacian matrix* $L(G)$ associated with (this labeling of) G is an n-by-n matrix. The (i,j) entry of $L(G)$ is $d(i)$, the degree of vertex i, when $i=j$; it is -1 when $\{i,j\}\in E$, and 0 otherwise. Two different labelings of G result in permutation similar Laplacian matrices. Indeed, G_1 and G_2 are isomorphic graphs if and only if $L(G_1)$ is permutation similar to $L(G_2)$. It follows that any similarity invariant of $L(G)$ is a property of the graph G, hence the extensive interest in "spectra of graphs". (See, e.g., [4].)

It was proposed in [15] to seek invariants which are more specific to permutation similarity. That paper initiated a serious study of the so-called Laplacian permanental polynomial, per $(xI-L(G))$. (Permanental polynomials of adjacency matrices had received attention earlier [4, pp. 34-36].) While the permanental polynomial is generically preserved only under monomial similarity (the good news), it is computationally intractable (the bad news) [18]. (In spite of this difficulty, Isabel Faria [5] has discovered a simple relationship between the multiplicity of 1 as a permanental root of $|L(G)|$, and the "star degree" of G. See § 5.)

It is the purpose of this paper to introduce a compromise between per and det. The second immanant of an n-by-n matrix $A=(a_{ij})$ is defined by

$$d_2(A)=\sum_{\sigma\in S_n}\chi_2(\sigma)\prod_{t=1}^{n}a_{t\sigma(t)},\tag{1}$$

where χ_2 is the irreducible character of S_n corresponding to the partition $(2,1^{n-2})$. In particular, $\chi_2(\sigma)=\varepsilon(\sigma)[F(\sigma)-1]$, where ε is the alternating character and F is the

* Received by the editors September 24, 1984, and in revised form November 1, 1985. This work was supported in part by the National Science Foundation under grant DMS-8300097.

† Department of Mathematics and Computer Science, California State University, Hayward, California 94542.

number of fixed points. Define the "d_2-polynomial" of G by

$$d_2(xI - L(G)) = c_0(G)x^n - c_1(G)x^{n-1} + \cdots + (-1)^n c_n(G). \tag{2}$$

Since d_2 is an immanant [10], the similarities which preserve the d_2-polynomial form a group which contains the invertible monomial matrices as a subgroup. As this subgroup is maximal in the (complex) general linear group (see [6] for a short proof), it follows that the d_2-polynomial is generically preserved only under monomial similarities (the good news). Moreover, if the time to compute the determinant of an n-by-n matrix is of the order n^3, the time to compute d_2 is of the order n^4 (more good news) [16].

There are some reasons to prefer the Laplacian matrix to the computationally simpler adjacency matrix. For one thing, $L(G)$ is positive semidefinite symmetric of rank $\leqq n-1$. (The rank of $L(G)$ is equal to $n-1$ if and only if G is connected.) Generally speaking, immanants match up well with positive semidefinite matrices (the value being the length of a "decomposable symmetrized tensor"). Moreover, d_2 is zero on a positive semidefinite matrix A if and only if A has a zero row (and column) or rank $A < n-1$ [12]. In addition, the greater complexity of $L(G)$, when compared with the adjacency matrix, suggests there may be fewer "algebraic accidents" [9].

2. Preliminary observations. First consider the d_2-polynomial of a general n-by-n matrix $A = (a_{ij})$. Denote by $Q_{k,n}$ the collection of nCk k-element subsets of $\{1, 2, \cdots, n\}$. For $X \in Q_{k,n}$, let $A[X]$ be the k-by-k principal submatrix of A corresponding to X. Denote by $A\{X\}$ the n-by-n matrix whose (i, j) entry is given by

$$A\{X\}_{ij} = \begin{cases} a_{ij} & \text{if } i \text{ and } j \in X, \\ \delta_{ij} & \text{otherwise,} \end{cases}$$

where δ_{ij} is the "Kronecker delta". Then $A\{X\}$ is (permutation similar to) the direct sum of $A[X]$ and the identity matrix of size $n-k$.

The coefficient, $c_k(A)$, of $(-1)^k x^{n-k}$ in the expansion of $d_2(xI - A)$ can be obtained in a manner which bears some similarity to the computation of its counterpart in the characteristic polynomial. Namely, it follows from (1) that

$$c_k(A) = \sum_{X \in Q_{k,n}} d_2 A\{X\}. \tag{3}$$

Denote by S_X the subgroup of S_m consisting of those permutations which individually fix all the integers not contained in X. Then S_X is naturally isomorphic to S_k (acting on the k elements of X). Since each element of S_X has $n-k$ more fixed points than the corresponding element of S_k, and since the corresponding permutations have the same signs, it follows from (1) and (3) that

$$c_k(A) = \sum_{X \in Q_{k,n}} (d_2 A[X] + (n-k) \det A[X]). \tag{4}$$

(Note that the "d_2" in (3) corresponds to a character of S_n and operates on n-by-n matrices, while the "d_2" in (4) corresponds to a character of S_k and operates on k-by-k matrices. This kind of complication does not arise with det or per because both ε and 1 remain irreducible upon restriction.)

Suppose $B = (b_{ij})$ is a k-by-k matrix, $k \geqq 2$. Since $d_2(B)$ weights the term corresponding to $\sigma \in S_n$ in det B by the factor $(F(\sigma) - 1)$, it follows that

$$d_2(B) = \sum_{i=1}^{k} b_{ii} \det B(i) - \det B, \tag{5}$$

where $B(i)$ is the submatrix of B obtained by deleting row and column i. Continuing from (4), we have (for $k \geqq 2$)

$$\begin{aligned} c_k(A) &= \sum_{X \in Q_{k,n}} \left(\sum_{i=1}^{k} a_{X_iX_i} \det A[X](i) + (n-k-1) \det A[X] \right) \\ &= \left(\sum_{X \in Q_{k,n}} \sum_{i=1}^{k} a_{X_iX_i} \det A[X](i) \right) + (n-k-1) q_k(A) \qquad (6) \\ &= \left(\sum_{X \in Q_{k-1,n}} \left(\sum_{t \notin X} a_{tt} \right) \det A[X] \right) + (n-k-1) q_k(A), \end{aligned}$$

where $q_k(A)$ is the coefficient of $(-1)^k x^{n-k}$ in the characteristic polynomial $\det(xI-A)$.

We now specialize to the case $A = L(G)$, rewriting (6) as

$$c_k(G) = (n-k-1)q_k(G) + \sum_{X \in Q_{k-1,n}} \left(\sum_{t \notin X} d(t) \right) \det L(G)[X], \tag{7}$$

where $q_k(G) = q_k(L(G))$. If G is a *regular* graph, then $d(t) = r$, say, $1 \leqq t \leqq n$, and we obtain

$$c_k(G) = (n-k-1)q_k(G) + (n-k+1) r q_{k-1}(G). \tag{8}$$

In this case, the d_2-polynomial affords no information not already contained in the characteristic polynomial of $L(G)$. We can say even more. Denote by $A(G)$ the adjacency matrix corresponding to G, $A(G) = D(G) - L(G)$, where $D(G)$ is the diagonal matrix of vertex degrees.

THEOREM 1. *Suppose G_1 and G_2 are two regular graphs. Then the following are equivalent*: (i) $\det(xI - L(G_1)) = \det(xI - L(G_2))$, (ii) $d_2(xI - L(G_1)) = d_2(xI - L(G_2))$, (iii) $\det(xI - A(G_1)) = \det(xI - A(G_2))$.

Proof. The equivalence of (i) and (ii) follows from (8) and the fact that (8) can be used to express the $q_k(G)$'s in terms of the $c_k(G)$'s. Assume (iii) holds. Since the degrees of the two polynomials are the same and the coefficients of x^{n-1} are the same, G_1 and G_2 have the same number of vertices and the same number of edges, and hence the same degree, r, of regularity. Substituting $(x-r)$ for x in $\det(xI + A(G_1)) = \det(xI + A(G_2))$ gives (i). Reversing the steps shows that (i) implies (iii). □

It follows from Theorem 1 and previously published work that there exist nonisomorphic graphs with the same (Laplacian) d_2-polynomial. For example, the two graphs in [4, Fig. 6.4] both afford the following polynomial (computed to 6 figures on the Hayward Cyber using the program in Appendix II):

$$\begin{aligned} &11x^{12} - 528x^{11} + 11400x^{10} - 146112x^9 + (1.23514\text{E}6)x^8 - (7.23264\text{E}6)x^7 \\ &+ (2.99626\text{E}7)x^6 - (8.79944\text{E}7)x^5 + (1.80372\text{E}8)x^4 - (2.48519\text{E}8)x^3 \\ &+ (2.13664\text{E}8)x^2 - (9.82057\text{E}7)x + (1.59252\text{E}7). \end{aligned}$$

We now abuse the language and *also* denote by $D(G)$ the *degree sequence* of G, i.e., $D(G) = (d(1), d(2), \cdots, d(n))$. If $e = \{i, j\} \in E$, let $D_e(G)$ be the *edge-deleted degree sequence* obtained from $D(G)$ by deleting $d(i)$ and $d(j)$. Denote by a_k the kth *elementary symmetric function* and define $a_k(G) = a_k(D(G))$. Finally, for $k \geqq 3$, let

$$b_k(G) = \sum_{e \in E} a_{k-2}(D_e(G)).$$

THEOREM 2. *Let G be a graph with n vertices and m edges. Denote by $L(G)$ the Laplacian matrix of G and write*

$$d_2(xI - L(G)) = \sum_{k=0}^{n} (-1)^k c_k(G) x^{n-k}.$$

Then

$$\begin{aligned}
&c_0(G) = n-1,\\
&c_1(G) = (n-1)a_1(G) = 2m(n-1),\\
&c_2(G) = (n-1)a_2(G) - m(n-3) \qquad (n \geqq 3),\\
&c_3(G) = (n-1)a_3(G) - (n-3)b_3(G) - 2(n-4)T(G) \qquad (n \geqq 4),
\end{aligned}$$

and

$$c_n(G) = d_2(L(G)) = 2m\kappa(G),$$

where $T(G)$ is the number of triangles (cycles of length 3) *in G, and $\kappa(G)$ is the number of spanning trees in G (i.e., the complexity of G).*

Proof. It follows from (1), with $A = xI - L(G)$, that $c_0(G) = \chi_2(id) =$ the degree of $\chi_2 = n-1$. The coefficient $c_1(G)$ is $\chi_2(id)$ times the trace of $L(G) = (n-1)a_1(G) = 2m(n-1)$. If $n \geqq 3$, contributions to x^{n-2} come from two sources. The first of these is $\sigma = id$, accounting for the $(n-1)a_2(G)$ term. The second source is the set of transpositions which interchange the vertices constituting an edge. In this case, the value of χ_2 is $-(n-3)$.

Contributions to $c_3(G)$ come from three sources: $(n-1)a_3(G)$ is contributed by $\sigma = id$, $-(n-3)b_3(G)$ arises from the transpositions which interchange the "endpoints" of an edge, and the remaining term comes from the 3-cycles $\sigma = (ijk)$ for which $\{i, j\}$, $\{j, k\}$ and $\{i, k\}$ are all edges. The value of χ_2 on a 3-cycle σ is $\varepsilon(\sigma)[F(\sigma) - 1] = 1[(n-3) - 1] = n-4$. But, the inverse of a 3-cycle is also a 3-cycle, i.e., each triangle in G makes both a clockwise and a counterclockwise contribution (giving at the office *and* at home). Finally, since the (i, j), (j, k), and (i, k) entries of $L(G)$, corresponding to the edges of the triangle, are all -1, the "contribution" is negative (perhaps explaining the previously noted generosity). In any event, the total contribution to $c_3(G)$ from the 3-cycles is $(-1)^3 2(n-4)T(G)$.

This brings us to $c_n(G)$. From (3), $c_n(G) = d_2(L(G))$. From (5), since $L(G)$ is singular,

$$c_n(G) = \sum_{i=1}^{n} d(i) \det(L(G)(i)).$$

But, from the "matrix-tree" theorem, $\det(L(G)(i)) = \kappa(G)$, for all i. □

Of course, the roots of $d_2(xI - L(G))$ are also graph-theoretic invariants. It is proved in [14] that these "d_2-roots" lie in the Gershgorin disks and in [11] that the *real* d_2-roots lie in the interval $[0, \lambda]$, where λ is the largest eigenvalue of $L(G)$.

We conclude this section with some inequalities for $c_k(G)$. The first of these is suggested by Theorem 2.

THEOREM 3. *Let G be a graph on n vertices. Denote by $a_k(G)$ the kth elementary symmetric function of $D(G)$ and by $c_k(G)$ the coefficient of $(-1)^k x^{n-k}$ in the polynomial $d_2(xI - L(G))$. Then $0 \leqq c_k(G) \leqq (n-1)a_k(G)$.*

Proof. Since $L(G)$ is positive semidefinite symmetric, all the terms in (7) are nonnegative. On the other hand, applying Hadamard's determinant inequality to the principal submatrices, (7) becomes

$$\begin{aligned} c_k(G) &\leqq (n-k-1)a_k(G) + \sum_{X\in Q_{k-1,n}} \left(\sum_{t\notin X} d(t)\right) \prod_{i\in X} d(i) \\ &= (n-k-1)a_k(G) + ka_k(G). \end{aligned} \qquad \square$$

THEOREM 4. *Let G be a graph on n vertices. Let G' be a spanning subgraph of G. Then $c_k(G) \geqq c_k(G')$, $k = 0, 1, \cdots, n$.*

Proof. Suppose $G = (V, E)$ and $G' = (V, E')$. Let $G_1 = (V, E\backslash E')$, so that G_1 is a graph on the same vertex set as G and G'. The set of edges of G_1 consists precisely of those edges of G which are not edges of G'. Then $L(G) = L(G') + L(G_1)$. If A and B are any two positive semidefinite hermitian matrices of the same size, $\det(A+B) \geqq \det A + \det B$. Therefore, the result follows from (7). □

COROLLARY 1. *Let G be a graph on n vertices. If $k \geqq 1$,*

$$c_k(G) \leqq \binom{n}{k} n^{k-2}[n(n-k)(n-1) + k(k-1)]. \tag{9}$$

Moreover, if G is connected, then $c_k(G) \geqq \max c_k(T)$, where the maximum is taken over all the spanning trees, T, of G. Finally, for connected G,

$$d_2(L(G)) = c_n(G) \geqq 2(n-1), \tag{10}$$

with equality if and only if G is a tree.

Proof. It follows from Theorem 4 that $c_k(G) \leqq c_k(K_n)$, where K_n is the complete graph. Since

$$q_k(K_n) = n^k \binom{n-1}{k},$$

it follows from (8) that

$$\begin{aligned} c_k(K_n) &= (n-k-1)n^k \binom{n-1}{k} + (n-k+1)(n-1)n^{k-1}\binom{n-1}{k-1} \\ &= \binom{n}{k} n^{k-2}[n(n-k)(n-1) + k(k-1)]. \end{aligned}$$

For a connected graph G, it follows from Theorem 4 that $c_k(G) \geqq c_k(T)$ for any one of its spanning trees, T. The characterization of trees in (10) is a consequence of Theorem 4 and the formula for $c_n(G)$ in Theorem 2. □

3. The coefficients $c_k(G)$. In this section, we obtain an explicit graph-theoretic characterization of the coefficients $c_k(G)$ in the (Laplacian) d_2-polynomial of G. We begin with (7). A. K. Kel'mans (see [2, Thm. 7.5] or [4, Thm. 1.4]) showed how to compute $q_k(G) = q_k(L(G))$. Consider all edge-subgraphs F of G which have k edges and are forests. Then

$$q_k(G) = \sum_F p(F), \tag{11}$$

where $p(F)$ is the product of the numbers of vertices in the connected components of F. To obtain a graph-theoretic characterization of $c_k(G)$, it remains to consider

$$\sum_{X\in Q_{k-1,n}} \left(\sum_{t\notin X} d(t)\right) \det L(G)[X]. \tag{12}$$

Following the treatment in [2], we make use of the fact that $L(G)=I(G)I(G)^t$, where $I(G)$ is the vertex-edge incidence matrix of G. Let $E=\{e_1, e_2, \cdots, e_m\}$ be the edges of G. Orient each edge (i.e., make it a directed edge) arbitrarily. Then $I(G)$ is the n-by-m matrix with (i,j) entry equal to $+1$ if vertex i is at the "positive" end of e_j, -1 if i is at the "negative" end of e_j, and 0 otherwise. In particular, each column of $I(G)$ contains exactly two nonzero entries. Then $L(G)=I(G)I(G)^t$, and this equation is independent of the orientation of the edges. It follows from the Cauchy-Binet theorem for the expansion of the determinant that

$$\text{(13)} \qquad \det L(G)[X] = \sum_{Y \in Q_{k-1,m}} \det I[X|Y]^2,$$

where $I[X|Y]$ is the $(k-1)$-square submatrix of $I(G)$ corresponding to the rows in X and columns in Y.

Poincaré proved that any square submatrix of $I(G)$ has determinant equal to 0, 1 or -1. Moreover [2, Lemma 7.4] $\det I[X|Y]$ is nonzero if and only if the edge-subgraph to which Y corresponds (which we will also call Y) is "*X-distinguished*" in the following sense:

DEFINITION 1. Let $G=(V,E)$ be a graph with n vertices and m edges. Suppose $n>t\geqq 1$ and let X be a t-element subset of V. (In our application, $t=k-1$.) Then the edge-subgraph $Y=(V_0, E_0)$ of G is *X-distinguished* if it satisfies the following four conditions:

(i) $o(E_0)=t$;

(ii) $X \subset V_0$;

(iii) Y is a forest;

(iv) $V_0 \backslash X$ contains exactly one vertex from each connected component of Y.

It follows from (13) and Poincaré's result that $\det L(G)[X]$ comprises a count of the *number of edge subgraphs of G which are X-distinguished.* Denote this number by $s(X)$, i.e.,

$$s(X)=\det L(G)[X].$$

Define

$$r(X)=\sum_{t\notin X} d(t).$$

Then we have proved the following result.

THEOREM 5. *Let $G=(V,E)$ be a graph with vertex set $V=\{1,2,\cdots,n\}$ and edge set $E=\{e_1,e_2,\cdots,e_m\}$. Denote by $L(G)$ the Laplacian matrix of G and write*

$$d_2(xI-L(G))=\sum_{k=0}^{n}(-1)^k c_k(G)x^{n-k}.$$

Then, for $k\geqq 2$,

$$\text{(14)} \qquad c_k(G)=(n-k-1)q_k(G)+\sum_{X\in Q_{k-1,n}} r(X)s(X),$$

where $q_k(G)$ is given in (11).

Example 1. We will implement (14) to calculate $c_3(G)$ for the graph in Fig. 1. The computation of $\sum r(X)s(X)$ is carried out in Table 1, where "$X=12$" represents $X=\{1,2\}$ and X-distinguished edge-subgraph 12 is $Y=(V_0,E_0)$, where $E_0=\{e_1,e_2\}$ and $V_0=\{1,2,6\}$ is the subset of V consisting of the "endpoints" of the edges in E_0. The computation of $q_3(G)$ is carried out in Table 2, where "$F=123$" denotes the edge-subgraph F with edge set $\{e_1,e_2,e_3\}$ (and vertex set $\{1,2,3,6\}$), a tree with

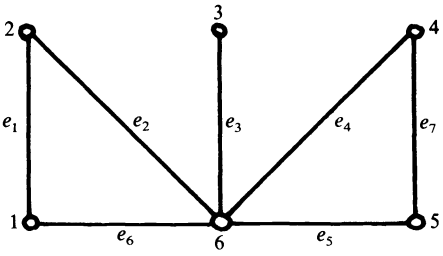


FIG. 1

4-vertices. Note $\sum p(F)=156$. Hence, from (14), $c_3(G)=592+2(156)=904$. (This is graph 19, $m=7$, in Appendix I.)

It is clear from Example 1 that Theorem 5 is not a useful computational device. Indeed (see Example 2 below) we can usually obtain $c_3(G)$ much more easily from Theorem 2. The point, of course, is that Theorem 5 is for general k while Theorem 2 involves only a few values of k. It remains an open problem to achieve a better

TABLE 1

X	X-distinguished edge-subgraphs	$s(X)$	$r(X)$	$s(X)r(X)$
12	12, 16, 26	3	10	30
13	13, 36	2	11	22
14	14, 17, 46, 67	4	10	40
15	15, 17, 56, 67	4	10	40
16	12, 13, 14, 15, 16, 26, 36, 46, 56	9	7	63
23	13, 23	2	11	22
24	14, 17, 24, 27	4	10	40
25	15, 17, 25, 27	4	10	40
26	12, 13, 14, 15, 16, 23, 24, 25, 26	9	7	63
34	34, 37	2	11	22
35	35, 37	2	11	22
36	23, 34, 35, 36	4	8	32
45	45, 47, 57	3	10	30
46	24, 27, 34, 37, 45, 46, 47, 57, 67	9	7	63
56	25, 27, 35, 37, 45, 47, 56, 57, 67	9	7	63
			total:	592

TABLE 2

F	$p(F)$	F	$p(F)$	F	$p(F)$
123	4	156	4	257	4
124	4	157	6	267	6
125	4	167	6	345	4
127	6	234	4	346	4
134	6	235	4	347	4
135	6	236	4	356	4
136	4	237	6	357	4
137	8	245	4	367	6
145	6	246	4	456	4
146	4	247	4	467	4
147	6	256	4	567	4

understanding, in general, of these graph-theoretic invariants. To some extent, this goal is accomplished for $c_{n-1}(G)$ in the next section.

Example 2. Let G be the graph in Fig. 1. We compute $c_3(G)$ by means of Theorem 2. So, $D(G)=(2,2,1,2,2,5)$, $a_3(G)=216$, $T(G)=2$, and

$$b_3(G)=\sum_{e\in E}\sum_{i\notin e} d(i)=\sum_{e\in E}\left(2m-\sum_{i\in e} d(i)\right)$$

$$=2m^2-\sum_{e\in E}\sum_{i\in e} d(i)=2m^2-\sum_{i=1}^{n} d(i)^2=2(7)^2-42=56.$$

Thus, $c_3(G)=5(216)-3(56)-2(2)(2)=904$.

We conclude this section by using Theorem 5 to give a second proof that $c_n(G)=2m\kappa(G)$. From (14),

$$c_n(G)=-q_n(G)+\sum_{X\in Q_{n-1,n}} r(X)s(X).$$

Since $L(G)$ is singular, $q_n(G)=\det L(G)=0$. Consider a fixed but arbitrary selection X of $n-1$ vertices, i.e., X consists of all but one of the vertices in V. We are looking for X-distinguished edge-subgraphs $Y=(V_0, E_0)$. Therefore $X\subset V_0$ and either $V_0=X$ or $V_0=V$. But, if $V_0=X$, then $o(V_0)=o(E_0)$ and Y cannot be a forest. Thus, $V_0=V$ and Y must be a tree. So, $s(X)=\kappa(G)$, while $r(X)$ is the degree of the vertex not contained in X. Finally,

$$\sum_{X\in Q_{n-1,n}} r(X)s(X)=\kappa(G)\sum_{X\in Q_{n-1,n}} r(X)$$

$$=\kappa(G)\sum_{i=1}^{n} d(i)=2m\kappa(G).$$

4. Vertex moments and centroids. In § 1, we argued that $c_k(G)$ is a "natural" graph-theoretic invariant, for so it seems from an algebraic point of view. Theorem 2 leaves open the possibility that $c_k(G)$ may even be (marginally) natural from the combinatorial viewpoint, but Theorem 5 might easily be viewed as closing the door on this possibility.

In this section, we use Theorem 5 to determine another (natural, even interesting) interpretation for $c_{n-1}(G)$. We begin with the special case that G is a tree.

DEFINITION 2. Let $G=(V,E)$ be a tree. For each $i,j\in V$, define $l(i,j)$ to be the *distance* from i to j, i.e., the length of the unique path in G from vertex i to vertex j. For each $i\in V$, the *moment* at i is

$$M(i)=\sum_{\substack{j\in V\\ j\neq i}} d(j)l(i,j).$$

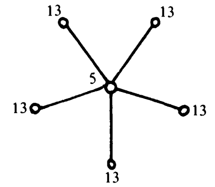
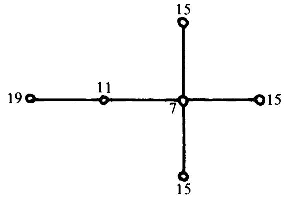
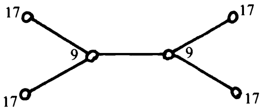
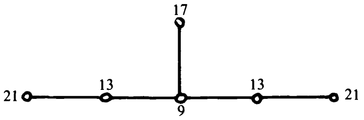
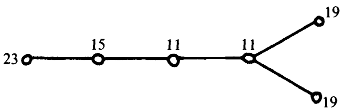
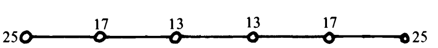
THEOREM 6. *Let $G=(V,E)$ be a tree with Laplacian matrix $L(G)$. Denote by $c_{n-1}(G)$ the coefficient of $(-1)^{n-1}x$ in the d_2-polynomial, $d_2(xI-L(G))$. Then*

$$c_{n-1}(G)=\sum_{i\in V} M(i).$$

This result is an immediate consequence of Theorem 8 below.

Example 3. Shown in Table 3 are the six trees on six vertices. Each vertex is labeled with its moment. In all cases, of course, the number of edges is 5. The symbol # refers to the number of the corresponding tree among the graphs on ($m=$) 5 edges in Appendix I.

Table 3

Graph	#	$c_5(G)$
	13	70
	9	82
	11	86
	10	94
	8	98
	7	110

Definition 3. Let $G=(V,E)$ be a tree with vertex set $V=\{1,2,\cdots,n\}$ and edge set E. Then vertex i is a *centroid point* of G if $M(i)=\min_{j\in V}M(j)$.

In Table 3, Graphs 9, 10 and 13 have unique centroid points while Graphs 7, 8 and 11 each have two (adjacent) centroid points.

It turns out that the name "centroid point" has already been preempted in the literature. (See, e.g., [7, Chap. 4].) A ***branch*** at vertex i of a tree is a maximal subtree containing i as an endpoint. (The number of branches at i is $d(i)$.) The ***weight*** at vertex i, $w(i)$, is the maximum number of edges in any branch at i. In the graph theory literature, i is a centroid point if $w(i)=\min_{j\in V}w(j)$. In the next result, we establish that these two definitions of "centroid point" are equivalent.

Theorem 7. *Let $G=(V,E)$ be a tree. Suppose $i\in V$. Then $M(i)=\min_{j\in V}M(j)$ if and only if $w(i)=\min_{j\in V}w(j)$.*

Proof. Each pair of adjacent vertices, i and j, of a tree divides the tree into two "intervals". To be precise, let $W(i,j)=\{v\in V|v\neq i$ but the unique path from v to j

passes through $i\}$. Define $\bar{W}(i,j)=W(i,j)\cup\{i\}$. Then $\bar{W}(i,j)$ is a branch at i and V is the disjoint union of $\bar{W}(i,j)$ and $\bar{W}(j,i)$.

Observe that

$$\begin{aligned} M(i) &= \sum_{\substack{v\in V\\ v\neq i}} d(v)l(i,v) \\ &= \sum_{v\in W(i,j)} d(v)l(i,v)+d(j)+\sum_{v\in W(j,i)} d(v)l(i,v), \\ M(j) &= \sum_{v\in W(i,j)} d(v)[l(i,v)+1]+d(i)+\sum_{v\in W(j,i)} d(v)[l(i,v)-1] \\ &= M(i)+\left[\sum_{v\in \bar{W}(i,j)} d(v)-\sum_{v\in \bar{W}(j,i)} d(v)\right]. \end{aligned}$$

So for adjacent vertices i and j,

$$M(j)-M(i)=\sum_{v\in \bar{W}(i,j)} d(v)-\sum_{v\in \bar{W}(j,i)} d(v). \tag{15}$$

Let $G(i,j)=(\bar{W}(i,j),\ E(i,j))$ be the subgraph of G with vertex set $\bar{W}(i,j)$ and edge set equal to the set of edges of G which join two vertices of $\bar{W}(i,j)$. Then

$$\sum_{v\in \bar{W}(i,j)} d(v)=2o(E(i,j))+1.$$

Thus,

$$M(j)-M(i)=2[o(E(i,j))-o(E(j,i))]. \tag{16}$$

It is known that there are at most two weight centroid points and that, if there are two, they are adjacent. Suppose, first, that there are two, say u_1 and u_2. Then $w(u_1)=1+o(E(u_2,u_1))$ and $w(u_2)=1+o(E(u_1,u_2))$. Since $w(u_1)=w(u_2)$, it follows from (16) that $M(u_1)=M(u_2)$. Let $v\notin\{u_1,u_2\}$. Without loss of generality, we may assume $v\in W(u_1,u_2)$. Then $w(v)>w(u_1)$. It suffices (in this case) to prove $M(v)> M(u_1)$. Let $i_0=u_1, i_1,\cdots, i_k=v$ be the successive vertices along the unique path from u_1 to v. Then $o(E(u_2,u_1))=w(u_1)-1<w(i_1)-1=o(E(u_1,i_1))$, i.e., the longest branch at i_1 is the one on which u_1 lies (not the one on which i_2 lies). Hence, $w(i_2)= o(E(i_1,i_2))+1,\cdots, w(v)=o(E(i_{k-1},v))+1$. Since $o(E(u_2,u_1))<o(E(u_1,i_1))< o(E(i_1,i_2))<\cdots<o(E(i_{k-1},v))$, it follows from (16) that both w and M are strictly, monotonically increasing from u_1 to v.

The proof in the case of a single weight centroid point is very similar. □

DEFINITION 4. Let $G=(V,E)$ be a graph with $n=o(V)>2$. Let $i,j\in V$ and take $X=\{v\in V|v\neq i \text{ and } v\neq j\}$. The *support of* $\{i,j\}$ *in* G is $s(i,j)=$ the number of X-distinguished edge-subgraphs of G.

Suppose, for example, that G is a tree on n vertices. Then $o(E)=n-1$. To compute the support of two vertices i and j, we count the number of X-distinguished subgraphs that can be obtained by deleting an edge of G. If the deleted edge does not lie on the path from i to j, then both i and j lie in the same component of the resulting subgraph and it is not X-distinguished ($V_0\backslash X\subset\{i,j\}$, with equality if the deleted edge does not separate i from j). If, on the other hand, the deleted edge does lie on the path from i to j, then the resulting subgraph is X-distinguished. Thus, for trees, $s(i,j)=l(i,j)$.

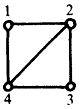
If G is a connected graph which is not a tree, consider one of its spanning trees T. Obtain subgraph T' by deleting one of the edges on the path from i to j in T. Call T' a *disconnection* of i and j. Then $s(i,j)=$ the total number of *different* disconnections of i and j.

DEFINITION 5. Let $G=(V, E)$ be a graph. For each $i \in V$, the *moment* of i is

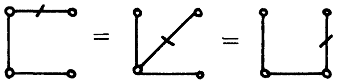
$$M(i)=\sum_{\substack{j \in V \\ j \neq i}} d(j)s(i,j). \tag{17}$$

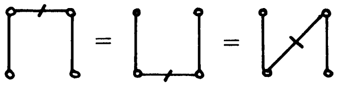
(Note that this reduces to Definition 2 when G is a tree.)

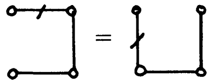
Example 4. Let

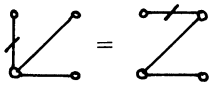
$G=$  .

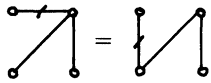
Then G has 8 spanning trees. For example, $s(1,2)=5$. The five $(1,2)$-disconnections are

(i) 

(ii) 

(iii) 

(iv) 

(v)  .

Of course, $s(1,2)=s(1,4)=s(2,3)=s(3,4)$. It can be shown that $s(2,4)=4$ and $s(1,3)=8$. Thus, $M(1)=3\cdot 5+2\cdot 8+3\cdot 5=46=M(3)$, while $M(2)=M(4)=2\cdot 5+3\cdot 4+2\cdot 5=32$.

THEOREM 8. *Let $G=(V, E)$ be a graph with Laplacian matrix $L(G)$. Denote by $c_{n-1}(G)$ the coefficient of $(-1)^{n-1}x$ in the d_2-polynomial, $d_2(xI-L(G))$. Then*

$$c_{n-1}(G)=\sum_{i \in V} M(i),$$

the moment sum of the vertices of G.

Proof. Suppose $V=\{1,2,\cdots,n\}$. From (14),

$$\begin{aligned} c_{n-1}(G) &= \sum_{X \in Q_{n-2,n}} r(X)s(X) \\ &= \sum_{i=1}^{n}\sum_{j=i+1}^{n} (d(i)+d(j))s(i,j) \\ &= \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \tfrac{1}{2}(d(i)+d(j))s(i,j) \\ &= \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} d(j)s(i,j)=\sum_{i=1}^{n} M(i). \end{aligned}$$

5. Some open problems. A. It is natural to want to extend the notion of centroid point to graphs G which are not trees. It is not obvious how to extend the weight definition. The moment definition, on the other hand, admits two candidates. A centroid point might be defined as a vertex with minimum moment in the sense of (17). On the other hand, at least for connected graphs, we might define centroid points to be those vertices i for which

$$\sum_{j\in V} d(j)f(i,j)$$

is minimized, where $f(i,j)$ is the minimum distance from i to j in G. How different are these two definitions when G is connected?

B. Let $R(G)=\{\rho_1,\rho_2,\cdots,\rho_n\}$ be the roots of $d_2(xI-L(G))$. As mentioned in § 2, $R(G)$ lies in the union of the Gershgorin disks, and the real ρ's lie in the interval $[0,\lambda]$ where λ is the largest eigenvalue of $L(G)$. Does $\operatorname{Re}(\rho_i)\in[0,\lambda]$ for $i=1,2,\cdots,n$? Faria [5] defined a *pendant star* to be a maximal subgraph formed by pendant edges all incident with the same vertex. The *degree* of a pendant star is one less than the number of its pendant edges. The *star degree* of a graph is the sum of the degrees of its pendant stars. Is the star degree of G always a lower bound for the multiplicity of 1 in $R(G)$? (It is for all graphs on 6 vertices. See Appendix I.) What is the exact relationship between $R(G)$ and $R(G')$, where G' is the dual of G? (For example, if the multiplicity of 1 in $R(G)$ is greater than one, does it always follow that $n-1\in R(G')$?)

C. Given n, m and k, can the graph G be characterized for which $c_k(G)$ is a maximum/minimum? For example, consider the set $T(n)$ of trees on n vertices. Let $S\in T(n)$ be the star (with a vertex of degree $n-1$) and $P\in T(n)$ be the path of length $n-1$. Is $c_{n-1}(S)\leqq c_{n-1}(T)\leqq c_{n-1}(P)$, for all $T\in T(n)$?

D. What is the "good" graph-theoretic description of $c_{n-2}(G)$?

E. Is $s(i,j)$ a "good" way to define a distance between vertices i and j?

Appendix I. Table 4 contains the coefficients of the d_2-polynomial for the graphs on $n=6$ vertices. If

$$d_2(xI-L(G))=\sum_{k=0}^{6}(-1)^k c_k(G)x^{n-k},$$

then $c_0(G)=5$ for all G and $c_1(G)=10m$ where m is the number of edges of G. Thus, these coefficients do not appear in the table. The symbol # indicates the graph number in Harary's tabulation [7, Appendix 1, $p=6$], and $D(G)$ is the degree sequence.

Appendix II. Equation (3) suggests that the computation of the d_2-polynomial, $d_2(xI-L(G))$, is a two step process. One generates $Q_{k,n}$ and then evaluates $d_2A\{X\}$. We do not claim any special insight into the generation of $Q_{k,n}$, but we do have a "fast" algorithm for evaluating d_2.

Using (1), the time to compute d_2 is clearly exponential in the size of the matrix. A similar problem arises in the case of the determinant. Of course, fast algorithms exist for determinant based (ultimately) on the fact that $\det(A)=0$ whenever A has two equal rows. Unfortunately, this "fact" is unavailable for any of the other immanants. (If A has *three* equal rows, $d_2(A)=0$.) Our algorithm uses (5) to convert the problem into one involving determinants.

What follows is a BASIC program to compute $d_2(xI-A)$, for any real, n-by-n matrix A, $3\leqq n\leqq 12$. Lines 5-65 are the preliminaries; lines 109-265 involve inputing the matrix. Lines 2000-2370 and 3050-3190 generate $Q_{k,n}$. Lines 2500-2650 set up $A\{X\}$. Lines 4300-5270 compute $d_2A\{X\}$ using (5) and row reduction. Line 3040 is the summation step in (5), and the output is produced in lines 275-460 and line 3200.

Table 4

m	#	$D(G)$	c_2	c_3	c_4	c_5	$c_6 = d_2$
0		0^6	0	0	0	0	0
1		$1^2, 0^4$	2	0	0	0	0
2	1	$2, 1^2, 0^3$	19	4	0	0	0
	2	$1^4, 0^2$	24	8	0	0	0
3	2	$2^3, 0^3$	51	18	0	0	0
	3	$3, 1^3, 0^2$	51	32	6	0	0
	1	$2^2, 1^2, 0^2$	56	36	6	0	0
	4	$2, 1^4, 0$	61	50	14	0	0
	5	1^6	66	64	24	0	0
4	4	$4, 1^4, 0$	98	104	49	8	0
	1	$3, 2^2, 1, 0^2$	103	94	24	0	0
	2	$2^4, 0^2$	108	112	32	0	0
	5	$3, 2, 1^3, 0$	108	122	57	8	0
	6	$2^3, 1^2, 0$	113	132	54	0	0
	3	$2^3, 1^2, 0$	113	136	65	8	0
	9	$3, 1^5$	113	146	88	20	0
	8	$2^2, 1^4$	118	160	98	20	0
	7	$2^2, 1^4$	118	160	101	24	0
5	13	$5, 1^5$	160	240	185	70	10
	5	$3^2, 2^2, 0^2$	170	220	80	0	0
	6	$4, 2^2, 1^2, 0$	170	244	149	30	0
	1	$3^2, 2, 1^2, 0$	175	258	157	30	0
	9	$4, 2, 1^4$	175	282	224	82	10
	3	$3, 2^3, 1, 0$	180	282	179	30	0
	4	$3, 2^3, 1, 0$	180	286	192	40	0
	11	$3^2, 1^4$	180	296	237	86	10
	2	$2^5, 0$	185	310	225	50	0
	15	$3, 2^2, 1^3$	185	316	250	72	0
	10	$3, 2^2, 1^3$	185	320	265	94	10
	8	$3, 2^2, 1^3$	185	320	268	98	10
	12	$2^4, 1^2$	190	340	285	90	0
	14	$2^4, 1^2$	190	344	296	96	0
	7	$2^4, 1^2$	190	344	301	110	10
6	1	$3^4, 0^2$	252	416	192	0	0
	15	$5, 2^2, 1^3$	252	488	471	216	36
	4	$4, 3, 2^2, 1, 0$	257	478	376	96	0
	5	$3^3, 2, 1, 0$	262	502	398	96	0
	3	$4, 2^4, 0$	262	512	429	108	0
	21	$4, 3, 2, 1^3$	262	526	515	232	36
	2	$3^2, 2^3, 0$	267	540	477	132	0
	6	$3^2, 2^3, 0$	267	544	492	144	0
	18	$3^3, 1^3$	267	550	543	240	36
	14	$4, 2^3, 1^2$	267	560	572	258	36
	12	$4, 2^3, 1^2$	267	564	592	288	48
	19	$3^2, 2^2, 1^2$	272	580	584	224	0
	20	$3^2, 2^2, 1^2$	272	584	607	276	36
	13	$3^2, 2^2, 1^2$	272	584	608	274	36
	11	$3^2, 2^2, 1^2$	272	588	625	300	48
	16	$3^2, 2^2, 1^2$	272	588	628	304	48
	10	$3, 2^4, 1$	277	618	678	318	36

TABLE 4 (continued)

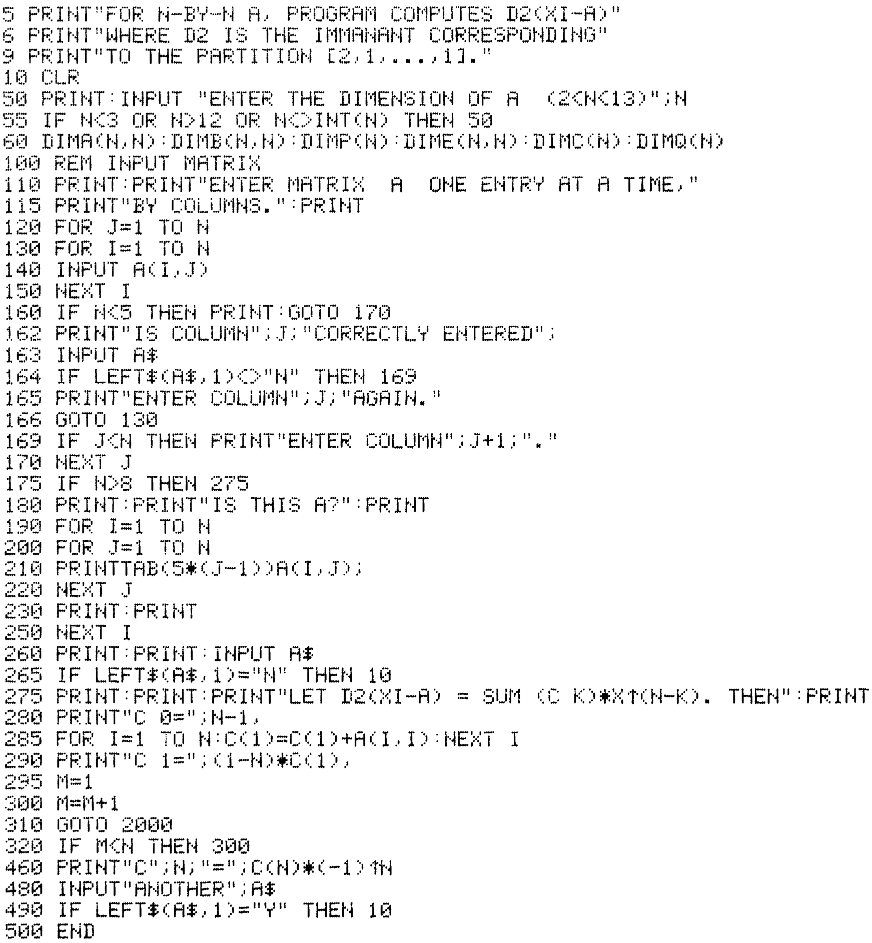
m	#	$D(G)$	c_2	c_3	c_4	c_5	$c_6 = d_2$
	9	$3, 2^4, 1$	277	622	696	344	48
	8	$3, 2^4, 1$	277	622	701	360	60
	17	2^6	282	648	729	324	0
	7	2^6	282	656	777	420	72
7	3	$4, 3^3, 1, 0$	359	812	760	224	0
	4	$4^2, 2^3, 0$	359	826	836	280	0
	18	$5, 3, 2^2, 1^2$	359	860	1018	564	112
	2	$4, 3^2, 2^2, 0$	364	860	889	294	0
	22	$4^2, 2^2, 1^2$	364	884	1054	580	112
	19	$5, 2^4, 1$	364	904	1121	642	126
	1	$3^4, 2, 0$	369	898	970	336	0
	24	$4, 3^2, 2, 1^2$	369	918	1110	608	112
	11	$4, 3^2, 2, 1^2$	369	918	1111	606	112
	15	$3^4, 1^2$	374	944	1120	512	0
	21	$3^4, 1^2$	374	952	1176	640	112
	8	$4, 3, 2^3, 1$	374	962	1226	692	112
	10	$4, 3, 2^3, 1$	374	962	1227	702	126
	20	$4, 3, 2^3, 1$	374	966	1252	750	154
	17	$4, 3, 2^3, 1$	374	970	1274	784	168
	9	$3^3, 2^2, 1$	379	996	1296	736	112
	12	$3^3, 2^2, 1$	379	1000	1322	794	154
	16	$3^3, 2^2, 1$	379	1000	1323	794	154
	14	$3^3, 2^2, 1$	379	1004	1342	824	168
	13	$4, 2^5$	379	1010	1372	860	168
	23	$3^2, 2^4$	384	1040	1421	858	126
	6	$3^2, 2^4$	384	1044	1453	938	196
	5	$3^2, 2^4$	384	1048	1473	970	210
	7	$3^2, 2^4$	384	1048	1478	988	224
8	18	$4^2, 3^2, 2, 0$	481	1328	1618	604	0
	4	$5, 3^3, 1^2$	481	1362	1868	1168	256
	3	$5, 4, 2^3, 1$	481	1376	1972	1328	320
	19	$4, 3^4, 0$	486	1376	1749	720	0
	22	$4^2, 3^2, 1^2$	486	1396	1934	1200	256
	2	$5, 3^2, 2^2, 1$	486	1420	2075	1408	336
	24	$4^2, 3, 2^2, 1$	491	1454	2144	1440	320
	11	$4^2, 3, 2^2, 1$	491	1454	2145	1452	336
	1	$5, 3, 2^4$	491	1474	2256	1616	384
	17	$4, 3^3, 2, 1$	496	1494	2228	1456	256
	20	$4, 3^3, 2, 1$	496	1498	2260	1538	336
	8	$4, 3^3, 2, 1$	496	1502	2288	1604	384
	10	$4, 3^3, 2, 1$	496	1502	2289	1604	384
	21	$4^2, 2^4$	496	1512	2364	1752	448
	15	$4^2, 2^4$	496	1520	2416	1856	512
	13	$3^5, 1$	501	1546	2408	1700	384
	14	$4, 3^2, 2^3$	501	1552	2450	1772	384
	12	$4, 3^2, 2^3$	501	1556	2484	1862	464
	16	$4, 3^2, 2^3$	501	1556	2485	1872	480
	9	$4, 3^2, 2^3$	501	1560	2512	1928	512
	5	$3^4, 2^2$	506	1600	2609	1984	480
	7	$3^4, 2^2$	506	1600	2614	2008	512
	6	$3^4, 2^2$	506	1604	2643	2076	560
	23	$3^4, 2^2$	506	1608	2665	2112	576

TABLE 4 (continued)

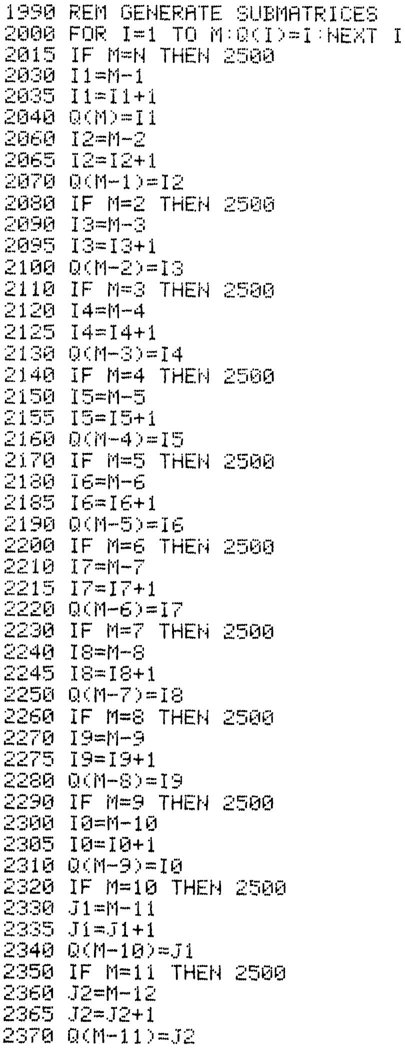
m	#	$D(G)$	c_2	c_3	c_4	c_5	$c_6 = d_2$
9	15	$4^3, 3^2, 0$	618	1984	2865	1350	0
	1	$5^2, 2^4$	618	2056	3504	2880	864
	4	$5, 4, 3^2, 2, 1$	623	2086	3484	2680	720
	21	$4^3, 3, 2, 1$	628	2130	3599	2766	720
	3	$5, 3^4, 1$	628	2144	3693	2934	810
	5	$5, 4, 3, 2^3$	628	2154	3788	3172	936
	12	$4^2, 3^3, 1$	633	2184	3778	2936	720
	14	$4^2, 3^3, 1$	633	2188	3812	3030	810
	18	$4^3, 2^3$	633	2198	3909	3294	972
	6	$5, 3^3, 2^2$	633	2204	3930	3264	864
	2	$5, 3^3, 2^2$	633	2208	3969	3382	990
	11	$4^2, 3^2, 2^2$	638	2252	4093	3494	990
	16	$4^2, 3^2, 2^2$	638	2252	4096	3512	1008
	20	$4^2, 3^2, 2^2$	638	2256	4129	3594	1080
	13	$4^2, 3^2, 2^2$	638	2256	4130	3606	1098
	19	$4^2, 3^2, 2^2$	638	2260	4160	3680	1152
	9	$4, 3^4, 2$	643	2310	4320	3840	1152
	8	$4, 3^4, 2$	643	2310	4325	3866	1188
	10	$4, 3^4, 2$	643	2314	4356	3942	1242
	7	3^6	648	2368	4557	4230	1350
	17	3^6	648	2376	4617	4374	1458
10	13	$4^5, 0$	770	2800	4625	2500	0
	6	$5, 4^2, 3^2, 1$	780	2980	5715	5000	1500
	5	$5^2, 3^2, 2^2$	780	3004	5952	5648	1920
	9	$4^4, 3, 1$	785	3034	5894	5170	1500
	1	$5, 4^2, 3, 2^2$	785	3058	6133	5860	1980
	11	$4^4, 2^2$	790	3112	6317	6060	2000
	4	$5, 4, 3^3, 2$	790	3122	6398	6236	2080
	3	$5, 4, 3^3, 2$	790	3126	6439	6368	2220
	10	$4^3, 3^2, 2$	795	3180	6629	6604	2280
	8	$4^3, 3^2, 2$	795	3180	6632	6622	2300
	15	$4^3, 3^2, 2$	795	3184	6668	6728	2400
	2	$5, 3^5$	795	3190	6715	6820	2420
	14	$4^2, 3^4$	800	3248	6944	7168	2560
	7	$4^2, 3^4$	800	3248	6949	7196	2600
	12	$4^2, 3^4$	800	3252	6987	7308	2700
11	4	$5, 4^4, 1$	952	4064	8725	8450	2750
	1	$5^2, 4, 3^2, 2$	957	4166	9418	10224	3960
	5	$5, 4^3, 3, 2$	962	4230	9683	10602	4070
	2	$5^2, 3^4$	962	4240	9784	10880	4224
	9	$4^5, 2$	967	4298	10000	11180	4400
	3	$5, 4^2, 3^3$	967	4308	10103	11482	4598
	6	$5, 4^2, 3^3$	967	4312	10146	11628	4752
	8	$4^4, 3^2$	972	4376	10422	12072	4928
	7	$4^4, 3^2$	972	4376	10425	12090	4950
12	3	$5^2, 4^3, 2$	1149	5548	13974	16860	7200
	2	$5^3, 3^3$	1149	5562	14148	17496	7776
	1	$5^2, 4^2, 3^2$	1154	5636	14514	18156	8064
	4	$5, 4^4, 3$	1159	5714	14936	19056	8640
	5	4^6	1164	5792	15360	19968	9216

TABLE 4 (continued)

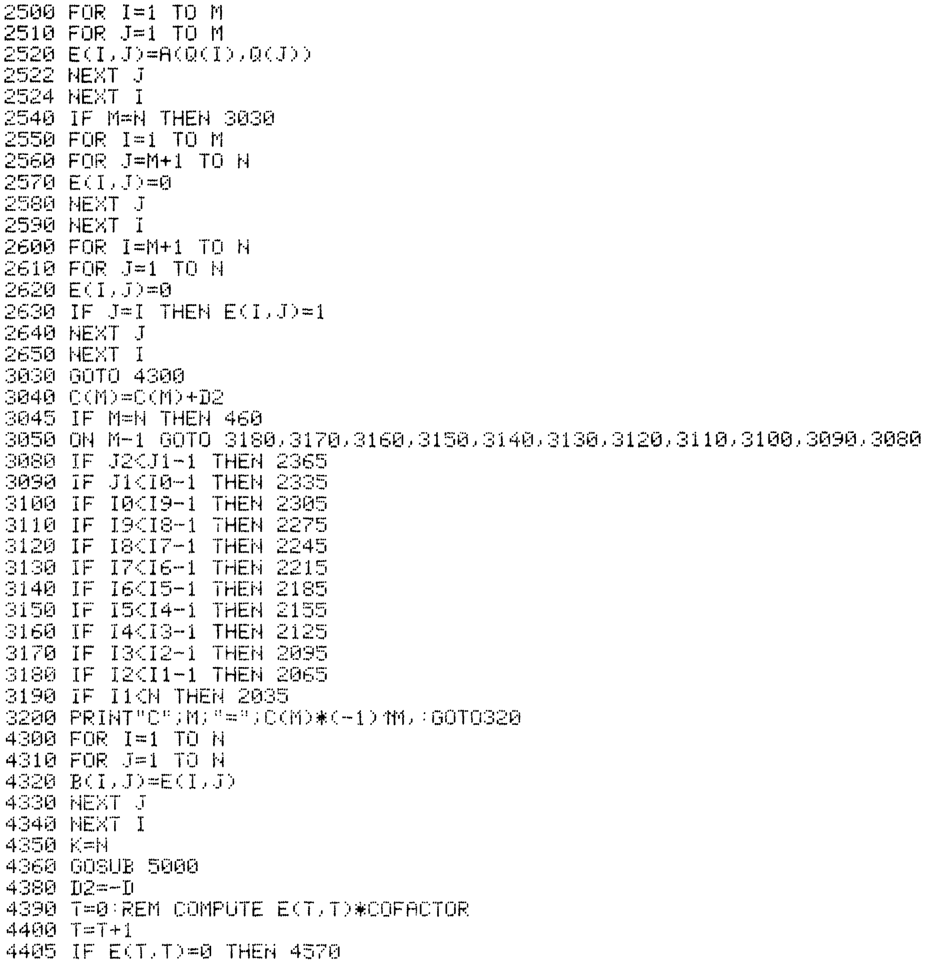
m	#	$D(G)$	c_2	c_3	c_4	c_5	$c_6 = d_2$
13	1	$5^3, 4^2, 3$	1361	7272	20586	28440	14040
	2	$5^2, 4^4$	1366	7360	21128	29760	14976
14		$5^4, 4^2$	1588	9256	28800	44064	24192
15		5^6	1830	11520	38880	64800	38880



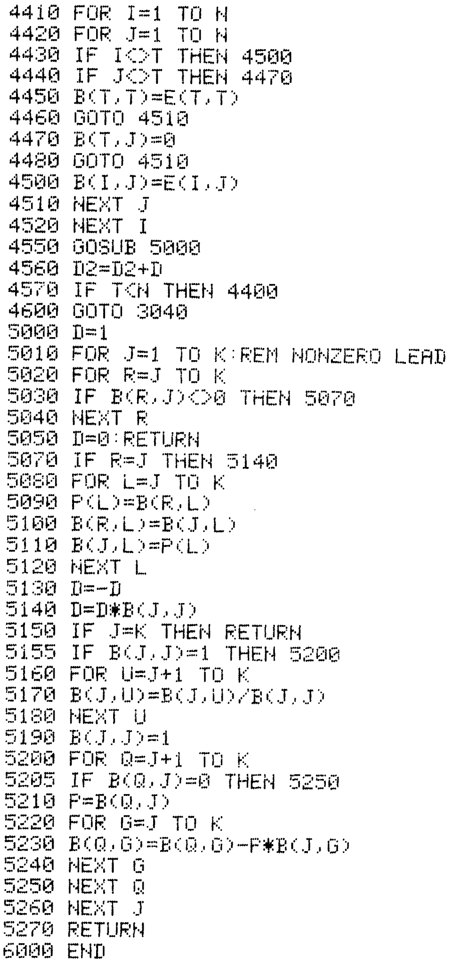
```
5 PRINT"FOR N-BY-N A, PROGRAM COMPUTES D2(XI-A)"
6 PRINT"WHERE D2 IS THE IMMANANT CORRESPONDING"
9 PRINT"TO THE PARTITION [2,1,...,1]."
10 CLR
50 PRINT:INPUT "ENTER THE DIMENSION OF A  (2<N<13)";N
55 IF N<3 OR N>12 OR N<>INT(N) THEN 50
60 DIMA(N,N):DIMB(N,N):DIMP(N):DIME(N,N):DIMC(N):DIMQ(N)
100 REM INPUT MATRIX
110 PRINT:PRINT"ENTER MATRIX  A  ONE ENTRY AT A TIME,"
115 PRINT"BY COLUMNS.":PRINT
120 FOR J=1 TO N
130 FOR I=1 TO N
140 INPUT A(I,J)
150 NEXT I
160 IF N<5 THEN PRINT:GOTO 170
162 PRINT"IS COLUMN";J;"CORRECTLY ENTERED";
163 INPUT A$
164 IF LEFT$(A$,1)<>"N" THEN 169
165 PRINT"ENTER COLUMN";J;"AGAIN."
166 GOTO 130
169 IF J<N THEN PRINT"ENTER COLUMN";J+1;"."
170 NEXT J
175 IF N>8 THEN 275
180 PRINT:PRINT"IS THIS A?":PRINT
190 FOR I=1 TO N
200 FOR J=1 TO N
210 PRINTTAB(5*(J-1))A(I,J);
220 NEXT J
230 PRINT:PRINT
250 NEXT I
260 PRINT:PRINT:INPUT A$
265 IF LEFT$(A$,1)="N" THEN 10
275 PRINT:PRINT:PRINT"LET D2(XI-A) = SUM (C K)*X↑(N-K). THEN":PRINT
280 PRINT"C 0=";N-1,
285 FOR I=1 TO N:C(1)=C(1)+A(I,I):NEXT I
290 PRINT"C 1=";(1-N)*C(1),
295 M=1
300 M=M+1
310 GOTO 2000
320 IF M<N THEN 300
460 PRINT"C";N;"=";C(N)*(-1)↑N
480 INPUT"ANOTHER";A$
490 IF LEFT$(A$,1)="Y" THEN 10
500 END
```



```
1990 REM GENERATE SUBMATRICES
2000 FOR I=1 TO M:Q(I)=I:NEXT I
2015 IF M=N THEN 2500
2030 I1=M-1
2035 I1=I1+1
2040 Q(M)=I1
2060 I2=M-2
2065 I2=I2+1
2070 Q(M-1)=I2
2080 IF M=2 THEN 2500
2090 I3=M-3
2095 I3=I3+1
2100 Q(M-2)=I3
2110 IF M=3 THEN 2500
2120 I4=M-4
2125 I4=I4+1
2130 Q(M-3)=I4
2140 IF M=4 THEN 2500
2150 I5=M-5
2155 I5=I5+1
2160 Q(M-4)=I5
2170 IF M=5 THEN 2500
2180 I6=M-6
2185 I6=I6+1
2190 Q(M-5)=I6
2200 IF M=6 THEN 2500
2210 I7=M-7
2215 I7=I7+1
2220 Q(M-6)=I7
2230 IF M=7 THEN 2500
2240 I8=M-8
2245 I8=I8+1
2250 Q(M-7)=I8
2260 IF M=8 THEN 2500
2270 I9=M-9
2275 I9=I9+1
2280 Q(M-8)=I9
2290 IF M=9 THEN 2500
2300 I0=M-10
2305 I0=I0+1
2310 Q(M-9)=I0
2320 IF M=10 THEN 2500
2330 J1=M-11
2335 J1=J1+1
2340 Q(M-10)=J1
2350 IF M=11 THEN 2500
2360 J2=M-12
2365 J2=J2+1
2370 Q(M-11)=J2
```



```
2500 FOR I=1 TO M
2510 FOR J=1 TO M
2520 E(I,J)=A(Q(I),Q(J))
2522 NEXT J
2524 NEXT I
2540 IF M=N THEN 3030
2550 FOR I=1 TO M
2560 FOR J=M+1 TO N
2570 E(I,J)=0
2580 NEXT J
2590 NEXT I
2600 FOR I=M+1 TO N
2610 FOR J=1 TO N
2620 E(I,J)=0
2630 IF J=I THEN E(I,J)=1
2640 NEXT J
2650 NEXT I
3030 GOTO 4300
3040 C(M)=C(M)+D2
3045 IF M=N THEN 460
3050 ON M-1 GOTO 3180,3170,3160,3150,3140,3130,3120,3110,3100,3090,3080
3080 IF J2<J1-1 THEN 2365
3090 IF J1<I0-1 THEN 2335
3100 IF I0<I9-1 THEN 2305
3110 IF I9<I8-1 THEN 2275
3120 IF I8<I7-1 THEN 2245
3130 IF I7<I6-1 THEN 2215
3140 IF I6<I5-1 THEN 2185
3150 IF I5<I4-1 THEN 2155
3160 IF I4<I3-1 THEN 2125
3170 IF I3<I2-1 THEN 2095
3180 IF I2<I1-1 THEN 2065
3190 IF I1<N THEN 2035
3200 PRINT"C";M;"=";C(M)*(-1)↑M,:GOTO320
4300 FOR I=1 TO N
4310 FOR J=1 TO N
4320 B(I,J)=E(I,J)
4330 NEXT J
4340 NEXT I
4350 K=N
4360 GOSUB 5000
4380 D2=-D
4390 T=0:REM COMPUTE E(T,T)*COFACTOR
4400 T=T+1
4405 IF E(T,T)=0 THEN 4570
```



```
4410 FOR I=1 TO N
4420 FOR J=1 TO N
4430 IF I<>T THEN 4500
4440 IF J<>T THEN 4470
4450 B(T,T)=E(T,T)
4460 GOTO 4510
4470 B(T,J)=0
4480 GOTO 4510
4500 B(I,J)=E(I,J)
4510 NEXT J
4520 NEXT I
4550 GOSUB 5000
4560 D2=D2+D
4570 IF T<N THEN 4400
4600 GOTO 3040
5000 D=1
5010 FOR J=1 TO K:REM NONZERO LEAD
5020 FOR R=J TO K
5030 IF B(R,J)<>0 THEN 5070
5040 NEXT R
5050 D=0:RETURN
5070 IF R=J THEN 5140
5080 FOR L=J TO K
5090 P(L)=B(R,L)
5100 B(R,L)=B(J,L)
5110 B(J,L)=P(L)
5120 NEXT L
5130 D=-D
5140 D=D*B(J,J)
5150 IF J=K THEN RETURN
5155 IF B(J,J)=1 THEN 5200
5160 FOR U=J+1 TO K
5170 B(J,U)=B(J,U)/B(J,J)
5180 NEXT U
5190 B(J,J)=1
5200 FOR Q=J+1 TO K
5205 IF B(Q,J)=0 THEN 5250
5210 P=B(Q,J)
5220 FOR G=J TO K
5230 B(Q,G)=B(Q,G)-P*B(J,G)
5240 NEXT G
5250 NEXT Q
5260 NEXT J
5270 RETURN
6000 END
```

Acknowledgment. The author is grateful for the careful, thorough job done by the referee, and for his/her many suggestions.

REFERENCES

[1] W. N. Anderson, Jr. and T. D. Morley, *Eigenvalues of the Laplacian matrix of a graph*, Linear and Multilinear Algebra, 18 (1985), pp. 141-145.

[2] N. Biggs, *Algebraic Graph Theory*, Cambridge Univ. Press, Cambridge, 1974.

[3] R. A. Brualdi and J. L. Goldwasser, *Permanent of the Laplacian matrix of trees and bipartite graphs*, Discrete Math., 48 (1984), pp. 1-21.

[4] D. M. Cvetković, M. Doob and H. Sachs, *Spectra of Graphs*, Academic Press, New York, 1980.

[5] I. Faria, *Permanental roots and star degree of a graph*, Linear Algebra Appl., 64 (1985), pp. 255-265.

[6] S. Friedland, *Maximality of the monomial group*, Linear and Multilinear Algebra, 18 (1985), pp. 1-7.

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

[8] A. J. Hoffman and D. K. Ray-Chandhuri, *On the line graph of a symmetric balanced incomplete block design*, Trans. Amer. Math. Soc., 116 (1965), pp. 238-252.

[9] C. R. Johnson and M. Newman, *A note on cospectral graphs*, J. Comb. Theory Ser. B, 28 (1980), pp. 96-103.

[10] D. E. Littlewood, *The Theory of Group Characters*, Oxford Univ. Press, Oxford, 1958.

[11] R. Merris, *Two problems involving Schur functions*, Linear Algebra Appl., 10 (1975), pp. 155-162.

[12] ———, *On vanishing decomposable symmetrized tensors*, Linear and Multilinear Algebra, 5 (1977), pp. 79-86.

[13] ———, *The Laplacian permanental polynomial for trees*, Czech. J. Math., 32 (107) (1982), pp. 397-403.

[14] C. R. Johnson, R. Merris and S. Pierce, *Inequalities involving immanants and diagonal products*, Portugaliae Math., to appear.

[15] R. Merris, K. R. Rebman and W. Watkins, *Permanental polynomials of graphs*, Linear Algebra Appl., 38 (1981), pp. 273-288.

[16] R. Merris and W. Watkins, *Inequalities and identities for generalized matrix functions*, Linear Algebra Appl., 64 (1985), pp. 223-242.

[17] J. Turner, *Generalized matrix functions and the graph isomorphism problem*, SIAM J. Appl. Math., 16 (1968), pp. 520-526.

[18] L. G. Valiant, *The complexity of computing the permanent*, Theoret. Comp. Sci., 8 (1979), pp. 189-201.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 4, October 1986

001

THE BANDWIDTH MINIMIZATION PROBLEM FOR CATERPILLARS WITH HAIR LENGTH 3 IS NP-COMPLETE*

BURKHARD MONIEN†

Abstract. It is shown that the *Bandwidth Minimization problem* remains NP-complete even when restricted to "caterpillars with hairs of length at most three". "Caterpillars" are special trees; they consist of a simple chain (the "body") with various simple chains attached to the vertices of the body (the attached chains are called "hairs"). A previous result in the literature shows that the bandwidth of caterpillars with hairs of length at most 2 can be found in $O(n \log n)$ time (this Journal, 2 (1981), pp. 387-393). We also show that the bandwidth problem is NP-complete when restricted to caterpillars with at most one hair attached to each vertex of the body. The proof is relatively straightforward and thereby also provides an easier proof than found in (SIAM J. Appl. Math., 34 (1978), pp. 477-495) that the bandwidth problem is NP-complete for trees with maximum vertex degree 3.

Key words. computational complexity, NP completeness, graph theory, bandwidth minimization

AMS(MOS) subject classifications. 68C25, 68E10

1. Introduction. An $n \times n$ matrix A is said to have *bandwidth* k if all of its nonzero entries are on one of the $2k+1$ diagonals consisting of the main diagonal and the k diagonals on either side of this main diagonal. The *Bandwidth Minimization problem* is to determine, for a given $n \times n$ matrix A and integer k, whether there exists an $n \times n$ permutation matrix P such that $P \cdot A \cdot P^T$ has bandwidth k. This problem is of great importance in many engineering applications. Typically, the matrices arising in these applications are sparse and matrix operations like inversion and multiplication can be performed with a considerably improved computation time if all the nonzero entries are placed within a small "band". Therefore the problem of reducing the bandwidth of a matrix has been of great interest during the last 20 years. A number of heuristics have been presented in the literature [1], [4], [7], [11]. The Bandwidth Minimization problem itself is NP-complete [13] implying (to our present knowledge) that there exists no efficient algorithm for solving this problem. The Bandwidth Minimization problem is equivalent to the following graph problem: given a graph G and an integer k, determine whether there exists a linear layout of G (i.e. integer labeling of the vertices of G such that each vertex receives a unique integer) such that the maximum difference between adjacent vertices is bounded by k. The problem has been studied also under a graph theoretic viewpoint [3], [4], [5], [6]. It is known to remain NP-complete even for trees with maximum vertex degree 3 [8].

On the positive side, dynamic programming algorithms have been described [10], [14] that can determine whether a graph G with n vertices has bandwidth k in at most $O(n^k)$ steps. It is also known that bandwidth 2 can be determined in linear time [8] and that there is a $O(n \log n)$ algorithm to determine the bandwidth of "caterpillars with hairs of length at most two" [2]. A "caterpillar" is a special kind of tree consisting of a simple chain C (called the "body" or "backbone") with an arbitrary number of simple chains attached by coalescing an endpoint of the added chain with a vertex in C. (The attached chains are called "hairs".) Caterpillars are shown in Fig. 1.1. A caterpillar has hairs of length at most k if all of the simple chains attached to the body have length at most k.

* Received by the editors June 12, 1985, and in revised form July 7, 1985.

† Fachbereich 17-Mathematik, Universität Paderborn, 4790 Paderborn, West Germany.

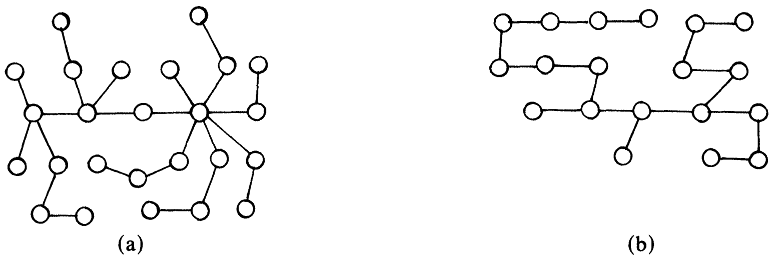


FIG. 1.1. (a) *A caterpillar with hairs of length at most* 3. (b) *A caterpillar with at most one hair attached to every vertex in the body.*

We show that the Bandwidth Minimization problem is NP-complete even when restricted to caterpillars with hairs of length at most 3 and that it is also NP-complete when restricted to caterpillars with at most one hair attached to every vertex in the body. Caterpillars of this latter type are a special kind of trees with maximum vertex degree 3. Our proof of NP-completeness is relatively straightforward and thereby provides an easier proof for the NP-completeness of the bandwidth problem on trees with maximum degree 3 [8].

In the case of caterpillars with at most one hair attached to any vertex of the body, we do not bound the length of the hairs. Caterpillars with maximum degree 3 and with hairs of length at most k have bandwidth at most k and therefore their Bandwidth Minimization problem can be decided in polynomial time [9], [13].

We have said above that the Bandwidth Minimization problem for caterpillars with hairs of length at most 2 is solvable in polynomial time whereas for caterpillars with hairs of length at most 3 it is NP-complete. The proof in § 2 will show that the border line we have determined is even sharper. We will see that the Bandwidth Minimization problem is NP-complete for caterpillars which have at most one node to which hairs of length 3 are attached, while all the other nodes of the body have hairs of length at most 1.

In [12] a weaker form of the NP-completeness result was shown. In the above paper a caterpillar is encoded as a chain with numbers attached to every node of the chain, i.e., the binary encoding is used for the length of the hairs. It is shown that with respect to this encoding the Bandwidth Minimization problem for caterpillars is NP-complete. In this interpretation a caterpillar is not viewed as a graph but as an instance of a special kind of scheduling problem. Note that under this encoding the length of the hairs may grow exponentially in the length of the encoding. In our paper we use the usual graph encoding.

We formulate the two results of this paper as theorems.

THEOREM 1. *The Bandwidth Minimization problem for caterpillars with hairs of length at most* 3 *is* NP-*complete.*

THEOREM 2. *The Bandwith Minimization problem for caterpillars with at most one hair attached to every vertex in the body is* NP-*complete.*

We will prove Theorem 1 in § 2 and Theorem 2 in § 3.

2. Caterpillars with hairs of length at most 3. We prove Theorem 1 by reduction from the Multiprocessor Scheduling problem [8, p. 238]. That is, given a set $T=\{t_1, t_2, \cdots, t_n\}$ of tasks (the ith task in T has execution time t_i), a deadline D, and a number m of processors, we construct a caterpillar C and an integer k such that C has bandwidth k if and only if the tasks in T can be scheduled on the m processors to satisfy the deadline D. The multiprocessor scheduling problem is strong NP-complete and therefore we can assume that all the t_i are polynomially bounded in n.

We first construct two portions of the caterpillar called "barrier" and "turning point". They are shown in Fig. 2.1.

The barrier of height p and the turning point of height p both have bandwidth p (a corresponding layout for the turning point of height 4 is shown in Fig. 2.2). Our construction of the caterpillar C is based on the fact that in every optimal layout of the turning point both nodes a and g either belong to the first half of the layout or to the second half of the layout, i.e., in every optimal layout of the turning point the backbone has to be folded. Because of the importance of this behaviour for our construction, we will give a careful proof below. Let T_p denote the turning point of height p. T_p has exactly $6p+1$ nodes.

LEMMA 1. *Let $T_p = (V, E)$, let $\sigma: V \to \{1, \cdots, 6p+1\}$ be a layout with $|\sigma(i)-\sigma(j)| \leqq p$ for all $\{i, j\} \in E$ and let $p \geqq 4$. Then either $\sigma(a), \sigma(g) < 3p+1$ or $\sigma(a), \sigma(g) > 3p+1$.*

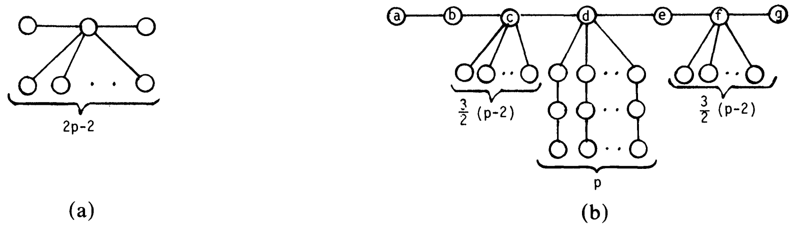
Proof. Let $v_0, \cdots, v_6$ be the nodes with $\sigma(v_i) = i \cdot p+1, 1 \leqq i \leqq 6$. We want to show first that $v_0 - v_1 - \cdots - v_6$ form a path in T_p and that $v_3 = d$ holds.

T_p is connected and every path in T_p has length at most 6. Since $\sigma(v_6) - \sigma(v_0) = 6p$ and since $|\sigma(i)-\sigma(j)| \leqq p$ for all $\{i, j\} \in E$, it follows that on the path from v_0 to v_6 any two adjcent nodes have the difference p with regard to σ. Therefore $v_0 - v_1 - \cdots - v_6$ is the only path from v_0 to v_6 of length 6. Every path of length 6 has the node d as its centre. This implies $d = v_3$.

We have seen that $\sigma(d) = 3p+1$ holds. This implies that σ can associate one of the numbers $1, \cdots, p$ or $5p+2, \cdots, 6p+1$ to a node u only if there exists a path of length at least 3 from d to u. This is true only if u is one of the nodes a, g or an endpoint of a hair of length 3 or a point on a hair dangling at the node f. Now let us assume that there exists an optimal layout σ with $\sigma(a) < \sigma(d) < \sigma(g)$. We have to show that this is not possible.

$\sigma(a) < \sigma(d)$ implies that not all the hairs of length 3 can be stretched out to the left. But then $\sigma(a) \leqq p$ must hold and $p-1$ hairs of length 3 together with the nodes a, b, c determine $3p$ nodes which have to be laid out to the left of d. The hairs of length 1 dangling at c and the two remaining neighbours of d have to get values from $[3p+2, 4p+1]$ with respect to σ. This is not possible since $\frac{3}{2}(p-2)+2 > p$ holds for $p \geqq 4$. □

The caterpillar C which we associate to the instance $Y = (\{t_1, \cdots, t_n\}, D, m)$ of the Multiprocessor Scheduling problem is shown in Fig. 2.3. We will see later that the number p has to fulfill some condition. We consider only instances Y with $\sum_{i=1}^{n} t_i = D \cdot m$. It is well known that the Multiprocessor Scheduling problem is strong NP-complete also when restricted to instances of this class.



(a) (b)

FIG. 2.1. (a) *The barrier of height p.* (b) *The turning point of height p, $p \equiv 0 \bmod 2$.*

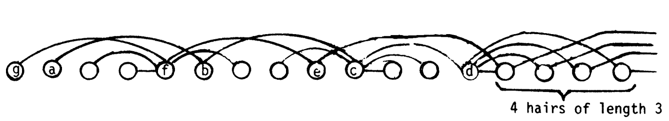


FIG. 2.2. *An optimal layout of the turning point of height* 4.

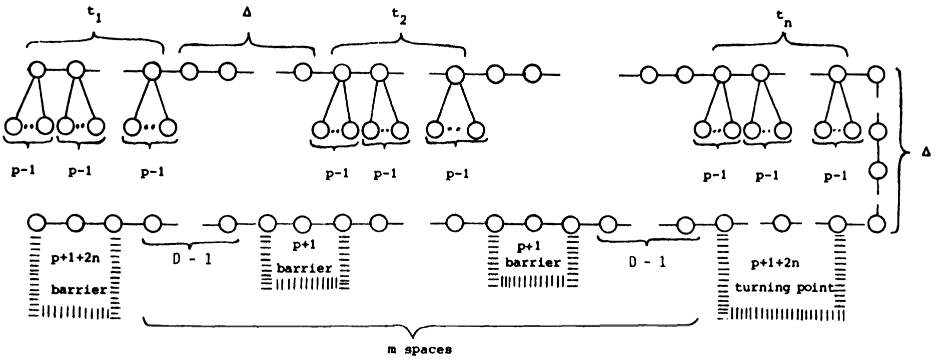


FIG. 2.3. *The caterpillar C, $\Delta = 2 \cdot \{m \cdot (D+2)-2\}$, p has to be chosen in an appropriate way.*

The caterpillar C consists of an encoding of the sequence of execution times and of some "frame" encoding m "holes" each of size D. These two parts are connected by the turning point of height $p+1+2n$. The behavior of the turning point will force every layout with bandwidth $p+1+2n$ to place these two parts one upon the other. It is a matter of technical details to show that a layout with bandwidth $p+1+2n$ exists if and only if each of the holes can be filled by using all the nodes of the blocks encoding certain execution times (this in turn is equivalent to the instance Υ having a solution).

We will give a formal proof in the remainder of this section. In order to make our description less cumbersome, we use some special terminology. The part of the body from the outermost barrier to the turning point we call "ground line" and the part from the turning point to the other end of the body we call "sweeping line". The "ith block", $1 \leqq i \leqq n$, consists of the chain of length t_i of the sweeping line together with the hairs attached to it. The ground line consists of $\lambda = m(D+2)+1$ nodes.

LEMMA 2. *If Υ has a solution then C has bandwidth $p+1+2n$.*

Proof. We will define the corresponding layout explicitly. Set $\beta = p+1+2n$.

(i) The ground line is stretched as far as possible, i.e. the points of the ground line get the numbers $i \cdot \beta + 1$, $1 \leqq i \leqq \lambda$. The hairs of the barriers are laid out in such a way that always half of them lie to the left of the center of the barrier and half of them to the right of the center.

(ii) For the turning point we use an optimal layout which associates with a and g the numbers $\lambda \cdot \beta + 1$ and $\lambda \cdot \beta + 2$. This layout uses the numbers $\lambda \cdot \beta + j$, $1 \leqq j \leqq 6 \cdot \beta + 1$, for the nodes of the turning point and has bandwidth β. All the remaining nodes (i.e. the nodes of the sweeping line and its hairs) get numbers smaller than $\lambda \cdot \beta + 1$.

(iii) Υ has a solution and therefore there exist sets I_j, $1 \leqq j \leqq m$, such that $\bigcup_{j=1}^{m} I_j = \{1, \cdots, n\}$ and $\sum_{i \in I_j} t_i = D$ for all $j = 1, \cdots, m$. For every i, if $i \in I_j$ then the nodes of the ith block are placed between the jth and the $(j+1)$st barrier (in the case $j = m$: between the mth barrier and the turning point). Note that by doing so we put exactly $D \cdot p$ nodes between any two barriers. Of course we must bear in mind that adjacent nodes have to get numbers at most β apart. But it is clear that this can be done and we can reach in this way a partial layout which fulfills the bandwidth constraint and which has the property that between any two nodes of the ground line there are laid out so far exactly p nodes.

(iv) Now we have to lay out the chains between the blocks. Every chain has length $\Delta = 2 \cdot (\lambda - 3)$. It can be laid out in such a way that the bandwidth constraint is fulfilled and that between any two nodes of the ground line there are placed exactly two nodes of the chain (see Fig. 2.4). That is, after we have laid out all the chains exactly $p+2n$ nodes are placed between any two nodes of the ground line, i.e., C has bandwidth $\beta=p+2n+1$. □

LEMMA 3. *If $p>2n\cdot(d+4)$ and if C has bandwidth $p+1+2n$, then Υ has a solution.*

Proof. We will show first that every layout σ of C with bandwidth $\beta=p+1+2n$ numbers the ground line up to symmetry in exactly the same way.

The turning point of height β is a subgraph of C. The turning point has $6\beta+1$ nodes and its longest path is of length 6. Therefore every layout with bandwidth β has to assign to its nodes $6\beta+1$ consecutive numbers. Because of Lemma 1 we know that the ground line and the sweeping line both lie with respect to σ either to the left of the turning point or to the right of the turning point. Therefore the nodes of the turning point have to get the smallest $6\beta+1$ numbers or the largest $6\beta+1$ numbers. We will assume that the turning point is laid out at the right end.

Furthermore the barrier of height β is a subgraph of C. This barrier has $2\beta+1$ nodes and its longest path is of length 2. Therefore σ associates to its nodes $2\beta+1$ consecutive numbers. No edge can cross this barrier and therefore we can conclude from the considerations made above that σ associates with the barrier of height β the numbers $1,\cdots,2\beta+1$. The current situation is shown in Fig. 2.5.

Note that C has

$$6\beta+1+(D-1)\cdot m+m\cdot(2p+3)+4n+n\cdot\Delta+p\cdot\sum_{i=1}^{n}t_i=\{6+m\cdot(D+2)\}\cdot\beta+1$$

nodes. We already know the numbers associated with the barrier of height β and the turning point. Between these two subgraphs the remaining $\{m\cdot(D+2)-2\}\cdot\beta-1$ nodes have to be laid out. The barrier of height β and the turning point are connected by the groundline, i.e. by a path of length $m\cdot(D+2)-2$. Therefore σ has to associate with the ith node of the ground line $1\leqq i\leqq\lambda$, the number $(i-1)\cdot\beta+1$.

We have seen that every layout with bandwidth β numbers the ground line and the turning point in the same way up to symmetry. We have to show now that the sweeping line can be encompassed into the "frame" given by the ground line and its barriers only if the scheduling problem Υ has a solution.

Note that the centers of the barriers have got the numbers $Z_j=\beta\cdot(D+2)\cdot j+\beta+1$, $j=0,\cdots m-1$. Set $Z_m=\beta(D+2)\cdot m+\beta+1$. We say that task i, $1\leqq i\leqq n$, belongs to the jth interval, $1\leqq j\leqq m$, if and only if $Z_{j-1}<\sigma(u)<Z_j$ holds for some node u of the sweeping line belonging to the ith block (i.e. to the subgraph of C encoding the execution time t_i). We will show first that a task cannot belong to two different intervals.

Let us assume that the task i belongs to two different intervals. Then there exist two adjacent nodes u, v belonging to the sweeping line and to the ith block such that

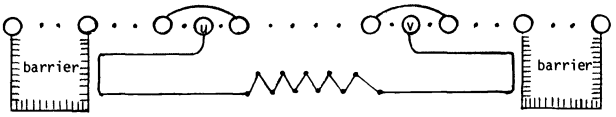


FIG. 2.4. *Layout of the chain connecting two blocks.*

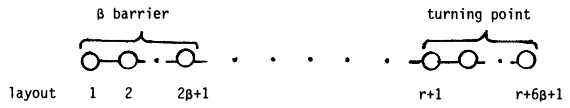


FIG. 2.5. *Layout σ, $r=m\cdot(D+2)\cdot\beta$.*

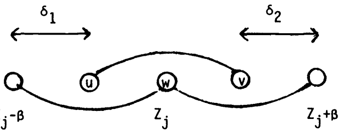


FIG. 2.6. *The situation where a task belongs to two intervals.*

$\sigma(u) < Z_j < \sigma(v)$ holds for some j. Let w be the node with $\sigma(w) = Z_j$. This situation is illustrated in Fig. 2.6. Set $\delta_1 = \sigma(u) - (Z_j - \beta)$ and $\delta_2 = Z_j + \beta - \sigma(v)$. $\{u, v\} \in E$ implies $\sigma(v) - \sigma(u) \leqq \beta$ and therefore $\delta_1 + \delta_2 \geqq \beta$ holds. There are $p-1$ hairs dangling at each of the nodes u and v and $2p$ hairs dangling at w. At most $\beta - \delta_1$ hairs dangling at u can get numbers smaller than $Z_j - \beta$ and at most $\beta - \delta_2$ hairs dangling at v can get numbers greater that $Z_j + \beta$. Therefore $5 + 2p + 2(p-1) - (\beta - \delta_1) - (\beta - \delta_2) \geqq 3 + 4p - \beta = 2 + 3p - 2n$ nodes have to get numbers between $Z_j - \beta$ and $Z_j + \beta$. This is not possible since $2\beta + 1 = 3 + 2p + 4n$ and $p \geqq 8n$ hold.

Thus we have shown that every task belongs to exactly one interval. Let I_j, $1 \leqq j \leqq m$ be the set of tasks belonging to the jth interval. We have to show that $\sum_{i \in I_j} t_i \leqq D$ holds for all $j = 1, \cdots, m$. σ has associated numbers between $Z_{j-1} - \beta$ and $Z_j + \beta$ to all the nodes belonging to a task from I_j (there are $p \cdot \sum_{i \in I_j} t_i$ such nodes) to the hairs of the two barriers ($4p$ nodes) and to the corresponding part of the groundline ($D + 5$ nodes). This implies

$$p \cdot \sum_{i \in I_j} t_i + 4p + D + 5 \leqq (D+4)(p + 2n + 1) + 1$$

and therefore

$$\sum_{i \in I_j} t_i \leqq D + \frac{2n(D+4)}{p}. \qquad \square$$

Theorem 1 follows from Lemma 2 and Lemma 3.

3. Caterpillars with at most one hair attached to every vertex in the body. The proof of Theorem 2 does not differ much from the proof of Theorem 1. This time we use a reduction from the 3-Partition problem [8, p. 224] which is a special case of the Multiprocessor Scheduling problem where only instances $(\{t_1, \cdots, t_n\}, D, m)$ with $n = 3m$ and $D/4 \leqq t_i \leqq D/2$, $1 \leqq i \leqq n$, are considered. We construct our caterpillar C again by using barriers and a turning point. The barriers and the turning point have to be defined now in a different way. They are shown in Figs. 3.1 and 3.2.

The barrier of height p has $(2p-1)^2$ nodes and the length of its body is equal to $4p-4$. It has bandwidth p (note that $p \cdot (4p-4) + 1 = (2p-1)^2$) and it is easy to see

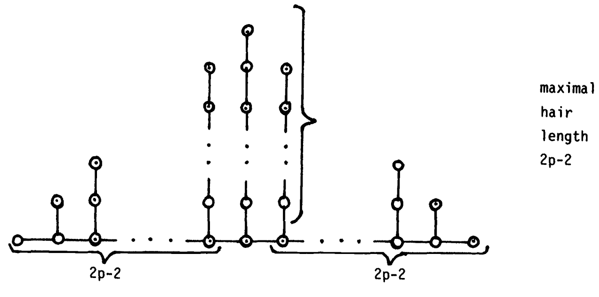


FIG. 3.1. *The barrier of height p.*

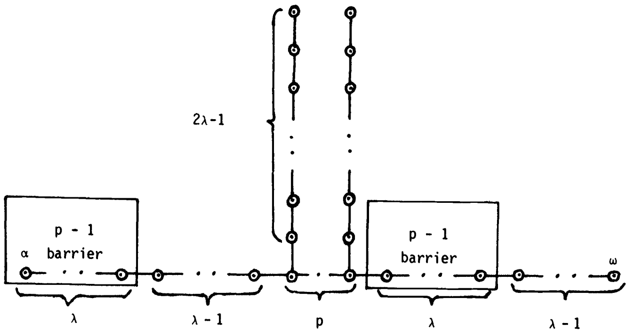


FIG. 3.2. *The turning point of height p, $\lambda = 4p-5$.*

that the strategy which lays out all hairs of the left part as far as possible to the left and all hairs of the right part as far as possible to the right leads to an optimal layout (in doing so we have to go from the outer hairs to the inner hairs and to pay attention to the bandwidth restriction). The turning point of the height p also has bandwidth p. This time we get an optimal layout by stretching all hairs of length $2\lambda - 1$ to the left and organizing the layout of the two barriers and the two chains such that both times one barrier and one chain overlap. Furthermore it is not difficult to see that for $p \geqq 5$ an analogue of Lemma 1 holds, i.e. for every layout with bandwidth p both nodes α and ω either belong to the left half of the layout or to the right half of the layout.

The caterpillar C which we associate this time to the instance $Y = (\{t_1, \cdots, t_n\}, D, m)$ of the 3-Partition problem is shown in Fig. 3.3. We can also apply the proof of Lemma 2 with only technical changes in this case showing that if Y has a solution then C has bandwidth $p+1+2n$.

In order to prove the other direction we follow the proof of Lemma 3. Again the ground line together with its barriers and its turning point defines a frame into which the sweeping line together with its hairs has to be embedded. As in the proof of Lemma 3 we define the notion of a task belonging to some interval.

A simple calculation shows that a task belongs to exactly one interval if $p \geqq 6n$ holds. Let us assume that a task i belongs to two intervals. Since the body of a barrier has length $4p+1$ and since nodes of the sweeping line belonging to task i are laid out as well to the left as to the right of the center of the barrier, there are more than p^2 nodes belonging to task i laid out within the region of the barrier. This is not possible if $p \geqq 6n$ holds.

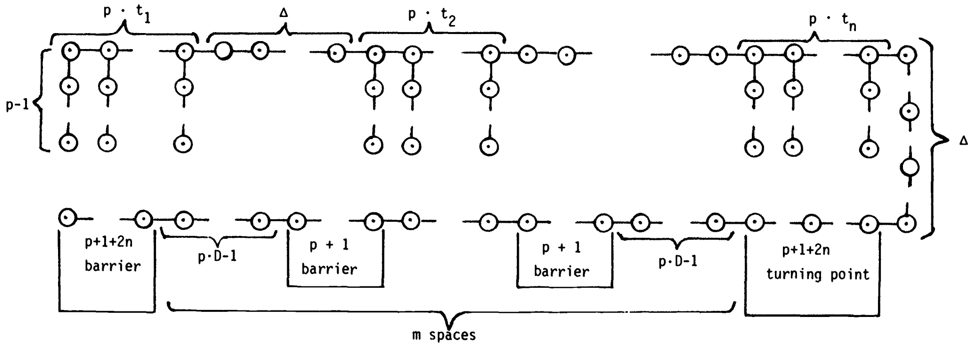


FIG. 3.3. *The caterpillar C, $\Delta = 2p\{m \cdot (D+4) - 4\}$, p has to be chosen in an approrpriate way.*

A similar straightforward computation shows that if I_j, $1 \leqq j \leqq m$, denotes the set of tasks belonging to the jth interval, then $\sum_{i \in I_j} t_i \leqq D$ holds for all $j = 1, \cdots, m$ provided p fulfills $p \geqq 2n \cdot (D+4)$. Thus we have shown that if C has bandwidth $p+1+2n$ and if $p \geqq 2n \cdot (D+4)$ holds then Υ has a solution. This completes the proof of Theorem 2. □

Acknowledgments. The author wants to thank I. H. Sudborough, O. Vornberger and R. Schulz for the numerous stimulating discussions we had during the preparation of this paper.

REFERENCES

[1] I. Arany, L. Szoda and W. F. Smyth, *An improved method for reducing the bandwidth of sparse symmetric matrices*, Proc. IFIP Conference 1971, North-Holland, Amsterdam, 1972, pp. 1246-1250.

[2] S. F. Assman, G. W. Peck, M. M. Syslo and J. Zak, *The bandwidth of caterpillars with hairs of length* 1 *and* 2, this Journal, 2 (1981), pp. 387-393.

[3] P. Z. Chinn, F. R. K. Chung, P. Erdös and R. L. Graham, *On the bandwidths of a graph and its complement*, in The Theory and Applications of Graphs, G. Chartrand, ed., John Wiley, New York, 1981, pp. 243-253.

[4] P. Z. Chinn, J. Chvátalová, A. K. Dewdney and M. B. Gibbs, *The bandwidth problem for graphs and matrices—A survey*, J. Graph Theory, 6 (1982), pp. 223-254.

[5] F. R. K. Chung, *Some problems and results on labelings of graphs*, in The Theory and Applications of Graphs, G. Chartrand, ed., John Wiley, New York, 1981, pp. 225-263.

[6] J. Chvátalová, *On the bandwidth problem for graphs*, Ph.D. thesis, Dept. Mathematics, Univ. Waterloo, Waterloo, Ontario, 1981.

[7] E. Cuthill and J. McKee, *Reducing the bandwidth of sparse symmetric matrices*, Proc. ACM National Conference 24, 1969, pp. 157-172.

[8] 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.

[9] 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.

[10] E. Gurari and I. H. Sudborough, *Improved dynamic programming algorithms for the bandwidth minimization problem and the min cut linear arrangement problem*, J. Algorithms, 5 (1984), pp. 531-546.

[11] W. Liu and A. H. Sherman, *Comparative analysis of the Cuthill-McKee ordering algorithms for sparse matrices*, SIAM J. Numer. Anal., 13 (1976), pp. 198-213.

[12] Z. Miller and J. Opatrny, *The complexity of the bandwidth problem for refinements of caterpillars*, unpublished paper.

[13] C. H. Papadimitriou, *The* NP-*completeness of the bandwidth minimization problem*, Computing, 16 (1976), pp. 237-267.

[14] J. B. Saxe, *Dynamic programming algorithms for recognizing small bandwidth graphs in polynomial time*, this Journal, 1 (1980), pp. 363-369.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 4, October 1986

002

A SPECTRUM ENVELOPING TECHNIQUE FOR ITERATIVE SOLUTION OF CENTRAL DIFFERENCE APPROXIMATIONS OF CONVECTION-DIFFUSION EQUATIONS*

MURLI M. GUPTA†

Abstract. When a convection-diffusion equation is discretized using the central difference scheme the resulting coefficient matrix is not diagonally dominant whenever the convection terms are large. If this system of linear equations is solved using the conventional iteration methods, the iterations often fail to converge as some of the eigenvalues of the iteration matrix lie outside the unit circle $C=\{z:|z|\leqq 1\}$ in the complex plane.

The eigenvalue spectrum of some of the iteration matrices lies inside the infinite strip $S=\{z:|\text{Real}(z)|<1, |\text{Imag}(z)|<\infty\}$. An example is that of the method of simultaneous displacements or the Jacobi method. In such cases, it is possible to enclose the eigenvalue spectrum inside an ellipse with major axis on the imaginary axis and minor axis in the real interval $(-1, 1)$. This ellipse is used to define a convergent iteration. A practical computational algorithm is described to obtain such an iteration scheme. Numerical examples show that the spectrum enveloping technique works well when the original iterations diverge. When the original iterations converge the spectrum enveloping technique can converge even faster.

Key words. spectrum enveloping, iterative methods, eigenvalues, central difference scheme, convection-diffusion equation, convection dominated flow

AMS(MOS) subject classifications. 65F10, 65N20, 76D05, 76-08.

1. Introduction.

Consider the convection-diffusion equation

$$\Delta u+R\frac{\partial u}{\partial x}=f(x, y), \tag{1}$$

where Δ represents the Laplacian operator and R represents the Reynolds number or the Peclét number.

When this differential equation is discretized at a mesh point (x, y) using central difference approximation, one obtains the following finite difference equation:

$$u_1+u_2+u_3+u_4-4u_0+(Rh/2)(u_1-u_3)=h^2 f_0, \tag{2}$$

where u_0 represents the value of u at (x, y) and the subscripts 1, 2, 3, 4 represent the four neighbouring values of u_0 at $(x+h, y)$, $(x, y+h)$, $(x-h, y)$ and $(x, y-h)$, respectively.

Equation (2) is defined on a set of K mesh points inside some two-dimensional domain Ω. Let w represent the vector of discrete values $u(x, y)$ at the K mesh points. The vector w is obtained from the linear system

$$Aw=b, \tag{3}$$

where A is a $K\times K$ coefficient matrix and b is a K vector.

The method of simultaneous displacements, or the Jacobi method, for solving (3) has the following form:

$$u_0^{(n+1)}=\frac{1}{4}\left[\left(1+\frac{Rh}{2}\right)u_1+u_2+\left(1-\frac{Rh}{2}\right)u_3+u_4\right]^{(n)}-\frac{h^2}{4}f_0. \tag{4}$$

It is well known that the Jacobi method is convergent when the coefficient matrix A is diagonally dominant [2], [7]. This happens when the mesh Reynolds number $|Rh/2|$

* Received by the editors January 10, 1985, and in revised form November 1, 1985.

† Department of Mathematics, The George Washington University, Washington, DC 20052. Currently at Mathematical Reviews, Ann Arbor, Michigan 48107.

is smaller than 1. In this case, other iteration techniques such as Gauss-Seidel and SOR are also convergent.

In this paper we consider the case when the mesh Reynolds number is larger than unity. This case represents the situation when the convection term $R\partial u/\partial x$ in (1) dominates the diffusion term Δu. In such cases, the coefficient matrix A loses its diagonal dominance and the conventional iteration methods fail to converge. This happens because some eigenvalues of the iteration matrix lie outside the unit circle $C=\{z: |z|\leqq 1\}$.

When the Jacobi iteration (4) is used to solve the linear system (3), the eigenvalue spectrum has the property that it lies inside an infinite strip $S=\{z: |\text{Real}(z)|<1, |\text{Imag}(z)|<\infty\}$; i.e., the real parts of all eigenvalues of the Jacobi iteration matrix lie in the interval $(-1, 1)$. In such a situation we can find an ellipse, in the complex plane with major axis on the imaginary axis and minor axis on the real interval $(-1, 1)$, which envelops the eigenvalue spectrum of the iteration matrix. If such an ellipse exists, then it is possible to define a new iteration method that is convergent for all values of the mesh Reynolds number. The new iteration also works when the basic iteration method used for solving (3) is convergent; in this case the new method can converge even faster.

The spectrum enveloping technique used in this paper was proposed by de Pillis [1]. Manteuffel [8] and Niethammer et al. [9], [10] give general results on elliptical enveloping of spectra. General results for K-step iterative methods and higher degree enveloping curves are given by Niethammer and Varga [10].

In the next section we examine the eigenvalue spectrum of Jacobi and other iteration matrices. In §§ 3 and 4, the spectrum enveloping technique is described and tested for Jacobi iterations. In § 5 we describe a practical algorithm for implementation of this procedure. This algorithm is then used on two test problems. In the last section we discuss the Gauss-Seidel method and its acceleration using the spectrum enveloping technique.

2. Eigenvalue spectrum of the iteration matrix. The coefficient matrix A of (3) can be split into $A=D+E+F$ where D is a diagonal matrix, E is a lower triangular matrix and F is an upper triangular matrix. The Jacobi iteration matrix is defined by [2], [7]:

$$J=-D^{-1}(E+F).$$

The successive over-relaxation (SOR) matrix is defined by

$$L_\omega=(D+\omega E)^{-1}(-\omega F+(1-\omega)D), \qquad 0<\omega<2.$$

A special case of the SOR matrix is the Gauss-Seidel matrix obtained with $\omega=1$:

$$L_1=-(D+E)^{-1}F.$$

The eigenvalue spectrum of these iteration matrices is denoted by $\{\zeta\}$, $\zeta=\xi+i\eta$.

When $R=0$ in (1), i.e., when the governing differential equation is the Poisson equation, all eigenvalues of the Jacobi matrix are real and are contained in $(-1, 1)$. The eigenvalue spectrum of the Jacobi matrix is also symmetric about the origin. The Gauss-Seidel matrix has real eigenvalues lying in $(0, 1)$. The spectrum of the SOR matrix depends upon the relaxation parameter ω and fills up the unit circle as the value of ω is increased.

The spectra of a test case are given in Fig. 1. This test case consists of a unit square $0\leqq x, y\leqq 1$ which is covered by a uniform $(N+1)\times(N+1)$ mesh (mesh width $h=N^{-1}$). A consistent ordering of mesh points is utilized to construct the coefficient matrix A

EIGENVALUES OF ITERATION MATRICES (R=0)

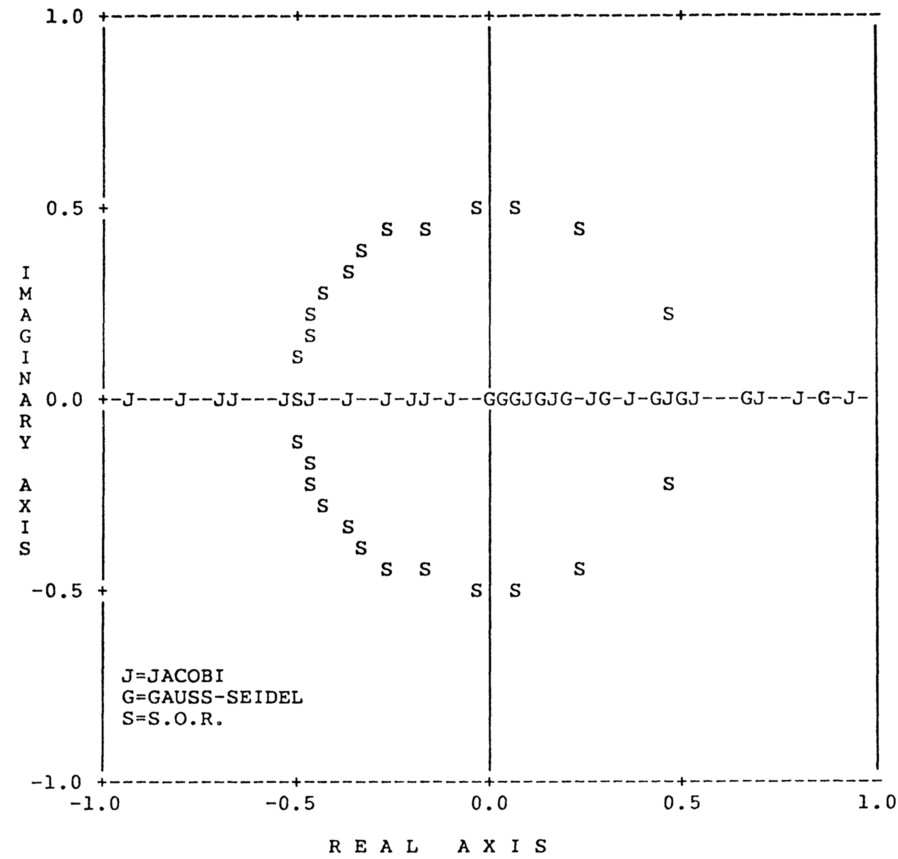


FIG. 1

[2], [7]. With a 9×9 mesh, the eigenvalues of the Jacobi iteration matrix (marked by J) lie on the real line between -0.9238 and 0.9238; the eigenvalues of the Gauss-Seidel matrix are marked by G; and the eigenvalues of the SOR matrix with $\omega=1.5$ are marked by S.

When the value of R is nonzero in (1), the Jacobi eigenvalues $\zeta(\zeta=\xi+i\eta)$ remain real and lie inside the interval $(-1,1)$ while $|Rh|<2$. When $|Rh|>2$, the eigenvalue spectrum spreads into the complex plane. The real part ξ is found to lie in the interval $(-1,1)$ and the maximum value of the imaginary part $|\eta|$ increases as the value of $|Rh|$ increases. In Figs. 2 and 3 we present the Jacobi eigenvalue spectrum for the 9×9 test case with $R=50$, 200 and 5000. It is noted that these spectra remain symmetric about both real and imaginary axes. Note also that the maximum value of $|\eta|$ is 1.3673 for $R=50$, 5.7552 for $R=200$ and 144.3361 for $R=5000$. In each case the real part ξ of every eigenvalue lies in the interval $(-.47,.47)$.

When the Gauss-Seidel method is used to solve the linear system (3) for $|Rh|>2$, the eigenvalue spectrum of the iteration matrix is found to lie mostly in the left half of the complex plane. In Figs. 4 and 5 we present such spectra for the test case when $R=50$ and $R=200$. We note that the maximum value of $|\xi|$ is 1.8703 with $R=50$ and 33.1702 with $R=200$. The maximum values of $|\eta|$ in these cases are 1.2634 and 5.2388

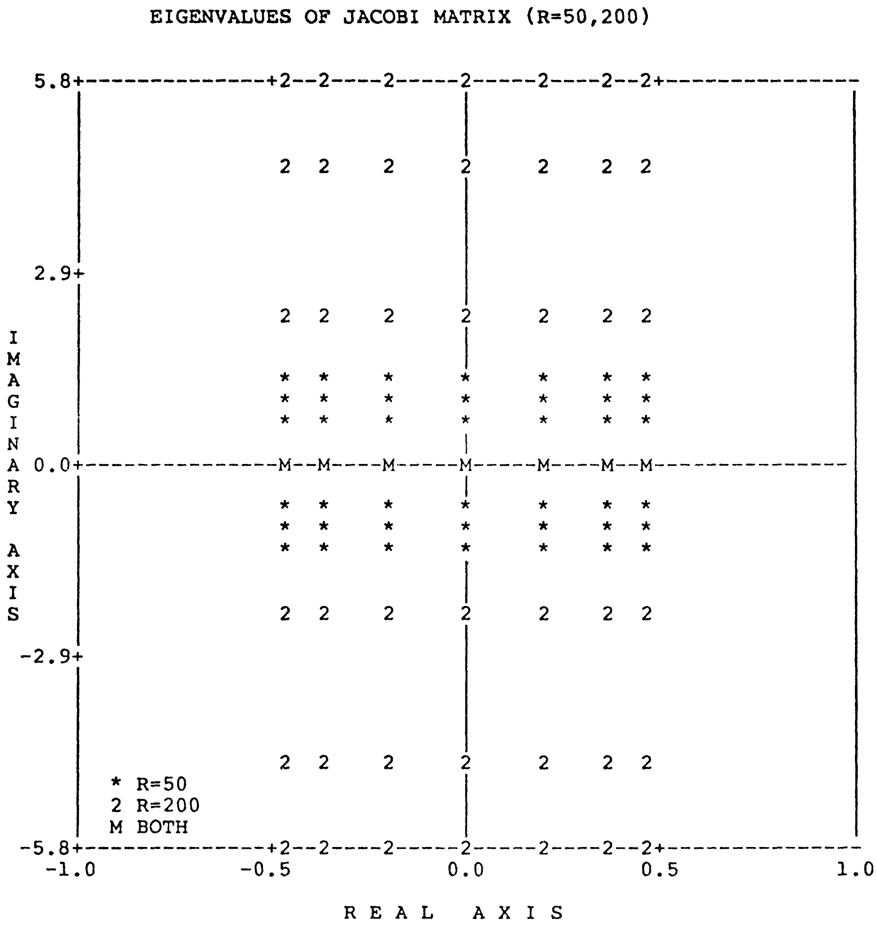


FIG. 2

respectively. It is observed that as $|Rh|$ increases, the Jacobi spectrum expands in both vertical directions keeping $|\xi|<1$ but the Gauss-Seidel spectrum expands mainly in the horizontal direction with the absolute value of the real part increasing as $|Rh|$ increases. The eigenvalue spectrum of the SOR matrix for $|Rh|>1$ with $\omega \neq 1$ looks similar to the Gauss-Seidel spectrum.

When the above iteration methods are used to solve the system of algebraic equations given by (2), the iterations generally converge when $|Rh| \leqq 2$ and diverge when $|Rh|>2$. In the next section we introduce a spectrum enveloping technique that converges even when the basic iterations diverge.

3. The spectrum enveloping technique. In a recent paper, de Pillis [1] proposed a technique to accelerate the convergence of an iteration scheme whose eigenvalue spectrum lies in the infinite vertical strip $S=\{z: |\text{Real}(z)|<1\}$. We describe this technique as follows:

A basic iteration method for solving $Aw=b$ is defined by

$$w_{n+1}=Bw_n+A_0^{-1}b \tag{5}$$

where the coefficient matrix A is written as $A=A_0(I-B)$ and A_0^{-1} is easy to find.

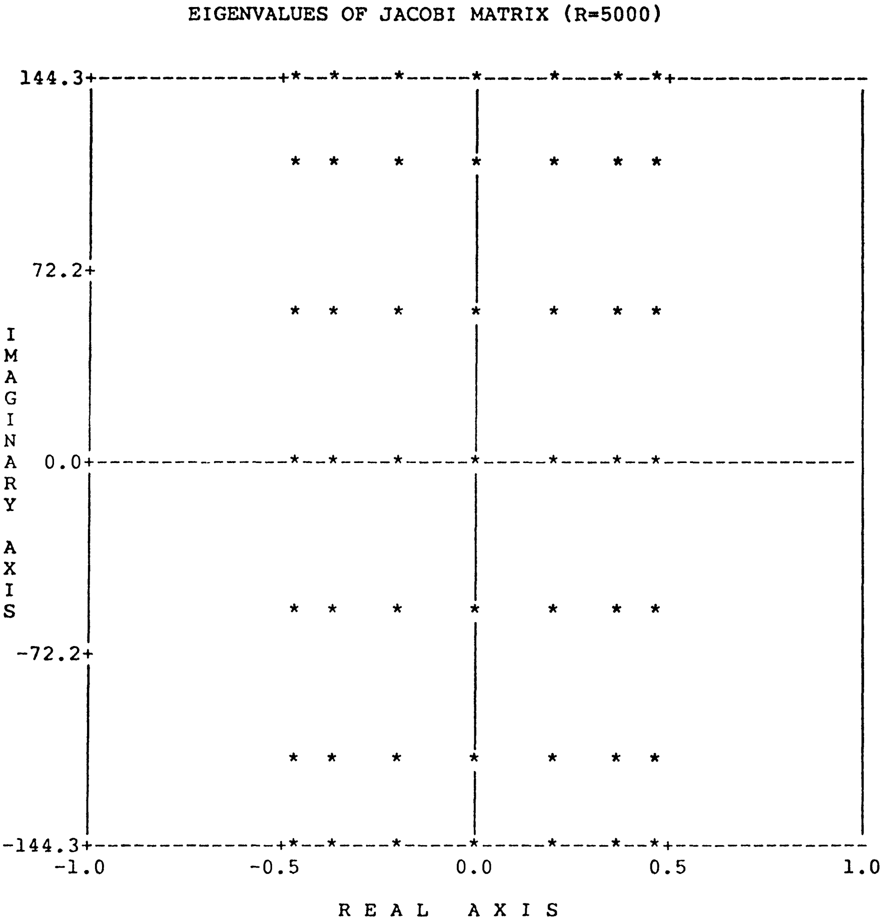


FIG. 3

The eigenvalue spectrum of the basic iteration matrix B is assumed to lie in the infinite vertical strip S. This spectrum can be enveloped (embraced in [1]) by a symmetric ellipse whose major semi-axis M lies on the imaginary axis and the minor semi-axis m lies on the real axis; $-1<m<1$.

Next define two constants λ and μ such that

$$\lambda=(m-M)/(m+M) \tag{6}$$

and μ is the unique root in $(0, 1)$ of the quadratic equation

$$(M+m)(1+\lambda\mu^2)=2\mu. \tag{7}$$

The new iteration scheme is defined as

$$y_{n+2}=(1+\lambda\mu^2)By_{n+1}-\lambda\mu^2y_n+(1+\lambda\mu^2)A_0^{-1}b. \tag{8}$$

De Pillis [1] showed that the sequence $\{y_n\}$ converges whenever $|\text{Real}\,(\zeta(B))|<1$. The asymptotic rate of convergence is given by $-\log_{10}\mu$. De Pillis also showed that the sequence $\{y_n\}$ converges faster than the sequence $\{w_n\}$ whenever the enveloping (embracing) ellipse has a nonzero eccentricity, i.e., $\lambda\neq 0$. In fact, the sequence $\{y_n\}$ converges even when the basic iteration scheme (5) is divergent assuming, of course, that the real part of all eigenvalues of the matrix B lies in the interval $(-1, 1)$.

EIGENVALUES OF GAUSS- SEIDEL MATRIX (R=50)

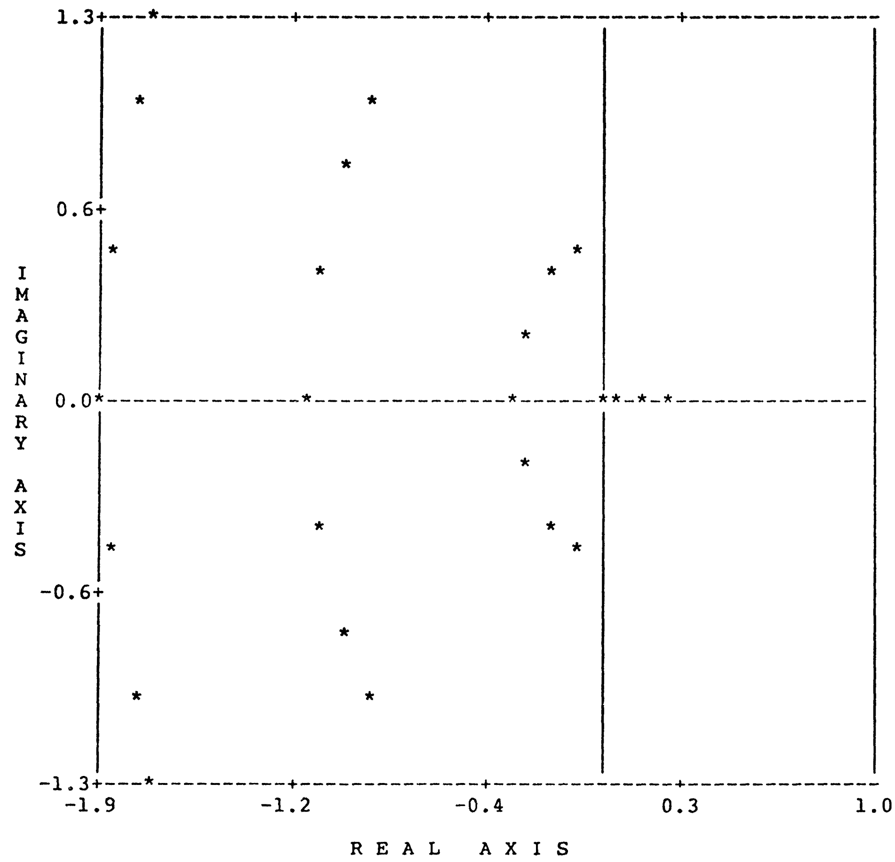


FIG. 4

Clearly, the Jacobi, the Gauss-Seidel and the Successive Over-Relaxation methods for solving the equation $Aw = b$ can be written in the form (5). The spectrum enveloping technique can be implemented through the following steps:

(1) Find the eigenvalue spectrum of the basic iteration matrix B.

(2) If the real part of all eigenvalues of B is in the interval $(-1, 1)$, find an enveloping ellipse that contains all eigenvalues of B.

(3) Use the semi-axes M, m of the enveloping ellipse to define λ, μ.

(4) Compute the solution of $Aw = b$ as follows:

Assume some initial approximations y_0, y_1. For $n = 0, 1, 2, \cdots$ compute using the basic iteration scheme an intermediate vector

$$\bar{y}_{n+2} = By_{n+1} + A_0^{-1}b. \tag{9}$$

Linear combination of $\bar{y}_{n+2}$ and y_n gives the new approximation:

$$y_{n+2} = (1 + \lambda\mu^2)\bar{y}_{n+2} - \lambda\mu^2 y_n. \tag{10}$$

It is clear that the combination of (9) and (10) is equivalent to (8).

4. Eigenvalue spectrum of the Jacobi iteration. When the central difference approximation (2) of (1) is solved by Jacobi iteration, the sequence $\{w_n\}$ is convergent only

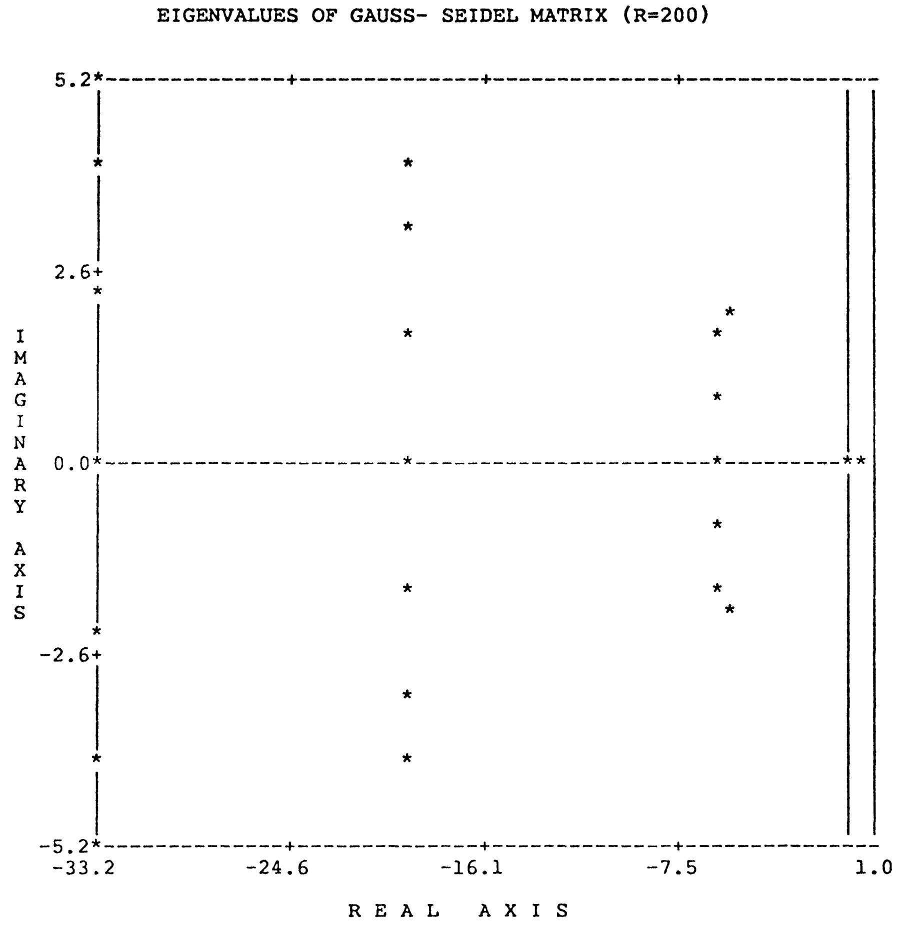


FIG. 5

when the value of $|Rh|$ is small. We consider the test problem 1 where the exact solution of (1) is given by $u(x, y) = xy(1-x)(1-y)$; the domain is the unit square $[0, 1] \times [0, 1]$ which is covered by a uniform $(N+1)\times(N+1)$ mesh $(N = h^{-1})$. The iterations are started from zero initial data and terminated when the maximum increment between the successive approximations is either smaller than 10^{-4} (convergence) or is larger than 100 (divergence). In Table 1, we give convergence and divergence data for the test problem 1.

In each case listed in Table 1, the Jacobi eigenvalues lie in the infinite strip S (see Figs. 2, 3 for the case $N=8$). In fact, $|\xi| < 0.5$. This means that the value of the semi-minor axis m (on the real axis) can be taken to be any number in the interval $(0.5, 1)$. We chose $m = 0.6$. The enveloping ellipse is given by

$$(x/m)^2 + (y/M)^2 = 1. \tag{11}$$

The value of M, the semi-major axis, is taken such that all eigenvalues of the iteration matrix lie inside the ellipse in (11): for any eigenvalue $\zeta(=\xi + i\eta)$ of the matrix B, $(\xi/m)^2 + (\eta/M)^2 < 1$. The value of M is chosen such that

$$M > m|\eta|(m^2 - \xi^2)^{-1/2}. \tag{12}$$

TABLE 1
Number of iterations for convergence (Jacobi iterations).

R	N			
	8	16	24	32
50	24*	41	58	102
100	10*	24*	119*	71
200	6*	10*	16*	25*
500	4*	6*	8*	9*
1000	3*	4*	5*	6*
2000	3*	3*	4*	5*
5000	2*	3*	3*	3*

(* = Divergence)

The semi-major axis M of the enveloping ellipse is such that (12) is satisfied for all complex eigenvalues ζ of the iteration matrix B.

In Table 2, we give the dominant eigenvalues of the 9×9 test case. An enveloping ellipse with $m=0.6$, $M=R/10$ satisfies (12) in each case. Corresponding values of λ and μ are obtained from (6) and (7). Since $\mu<1$, the iteration scheme defined in (9), (10) is convergent. In Table 2 we also give the number of iterations required for a 10^{-4} convergence. It is noted that the spectrum enveloping method converges for all values of R whereas the standard Jacobi diverged for all values of $R\geqq 50$ $(N=8)$ as seen in Table 1.

TABLE 2
Dominant eigenvalue and spectrum enveloping data $(N=8,\ m=0.6,\ M=R/10)$.

R	Dominant eigenvalue ξ	η	λ	μ	Number of iterations
50	.4619	1.3673	−.8182	.9054	49
100	.4619	2.8495	−.9048	.9513	89
200	.4619	5.7552	−.9512	.9753	187
500	.4619	14.4265	−.9802	.9901	549
1000	.4620	28.8635	−.9900	.9950	1211
2000	.4619	57.7343	−.9950	.9975	2622
5000	.4619	144.3361	−.9980	.9990	7419

5. A practical algorithm. In a practical situation, one does not know, a priori, the location of the eigenvalue spectrum of the iteration matrix. In order to define the enveloping ellipse, one must first locate the largest eigenvalue of the basic iteration matrix B. If these eigenvalues lie in the infinite strip $S=\{z\colon |\mathrm{Real}\,(z)|<1\}$, then the method of this paper would be applicable. If the eigenvalue spectrum of B lies outside the set S, the spectrum enveloping technique could not be used.

We now describe a practical algorithm to locate the eigenvalue spectrum, to define the enveloping ellipse and to carry out the new iterations. This algorithm consists of three phases:

5.1. Phase 1: eigenvalue estimation.

(i) Carry out a small number of basic iterations (even if this process is divergent):

$$\begin{gathered}\text{Initial vector } w_0\\ w_{n+1}=Bw_n+A_0^{-1}b\end{gathered}\tag{13}$$

(we prescribed a maximum of 20 iterations).

(ii) Begin estimating the dominant eigenvalues of B: At the nth iteration,

(a) Compute successive ratios

$$r_n=\|w_{n+1}-w_n\|/\|w_n-w_{n-1}\|,\qquad n=2,3,\cdots.\tag{14}$$

Sometimes alternate ratios might be used:

$$s_n=(\|w_{n+1}-w_n\|/\|w_{n-1}-w_{n-2}\|)^{1/2}\tag{15}$$

(we used the maximum norm).

(b) Continue iterations (13) until the ratios r_n or the ratios s_n settle down, i.e.,

$$|r_n-r_{n-1}|<\delta\quad\text{or}\quad|s_n-s_{n-1}|<\delta\tag{16}$$

for some tolerance δ (we used $\delta=.05$).

(c) When the ratios (14) or (15) have settled down, take the value of successive increment vectors

$$w_{n+1}-w_n,\quad w_n-w_{n-1},\quad w_{n-1}-w_{n-2}$$

at any two mesh points P and Q and define $y_1, y_2, y_3, z_1, z_2, z_3$ as follows:

$$\begin{aligned}y_3&=(w_{n+1}-w_n)(P), & z_3&=(w_{n+1}-w_n)(Q),\\ y_2&=(w_n-w_{n-1})(P), & z_2&=(w_n-w_{n-1})(Q),\\ y_1&=(w_{n-1}-w_{n-2})(P), & z_1&=(w_{n-1}-w_{n-2})(Q)\end{aligned}\tag{17}$$

(we took P to be the point where $|w_{n+1}-w_n|$ is largest; Q was one of the neighbours).

(d) Compute den $=y_2z_1-y_1z_2$.
If den $=0$, go to step (i).
If den $\neq 0$, compute

$$p=(z_3y_1-y_3z_1)/\text{den},\qquad q=(z_2y_3-y_2z_3)/\text{den}.\tag{18}$$

(e) The dominant eigenvalues are roots of

$$\zeta^2+p\zeta+q=0.\tag{19}$$

Compute $\Delta=p^2-4q$.
If $\Delta\geqq 0$ (real roots), $\xi=[-p\pm\sqrt{\Delta}]/2$, $\eta=0$.
If $\Delta<0$ (complex roots), $\xi=-p/2$, $\eta=\sqrt{-\Delta/2}$.

(f) Test for acceptability of ζ.
Compute $|\zeta|=\sqrt{(\xi^2+\eta^2)}$.
If $|r_n-|\zeta||/r_n\leqq\varepsilon$ (acceptable), go to Phase 2.
If $|r_n-|\zeta||/r_n>\varepsilon$ or if $|\xi|>0.995$, go to step (i)
(we used $\varepsilon=0.1$).

(iii) If a suitable eigenvalue is not estimated in steps (i), (ii) in a preassigned number of iteration, the computation is terminated.

5.2. Phase 2: spectrum enveloping. When the dominant eigenvalue ζ has been computed, the enveloping ellipse is defined as follows:

(i) Define the semi-minor axis m such that $m > |\xi|$. (We chose

$$m = \begin{cases} |\xi| + .05 & \text{if } |\xi| < .9, \\ (|\xi|+1)/2 & \text{if } |\xi| \geqq .9, \end{cases} \tag{20}$$

with the provision that $m \geqq 0.6$.)

(ii) Define the semi-major axis M from (12)

$$M = m|\eta|(m^2 - |\xi|^2)^{-1/2},$$

with the provision that $M \geqq 1$.

(iii) Compute $\lambda = (m - M)/(m + M)$;

$$\mu = [1 - \sqrt{\{1 - \lambda(m+M)^2\}}]/[\lambda(m+M)].$$

5.3. Phase 3: convergent iterations.

(i) Define initial vectors y_0, y_1.

(ii) Compute the numerical solution from (8):

$$y_{n+2} = (1 + \lambda\mu^2)(By_{n+1} + A_0^{-1}b) - \lambda\mu^2 y_n.$$

6. Numerical examples. The first test problem is (1) with the exact solution $u(x, y) = xy(1-x)(1-y)$ in the unit square $[0, 1] \times [0, 1]$. The right-hand function is given by $f(x, y) = -2[x(1-x) + y(1-y)] + R(1-2x)y(1-y)$. As noted in Table 1, the standard Jacobi iteration for solving the resulting system of algebraic equations diverges whenever $|Rh|$ is large. Using the algorithm outlined in the last section, we estimated the dominant eigenvalues of the Jacobi iterations for $N = 8, 16, 24, 32$ ($N = h^{-1}$). Once a satisfactory estimate of ζ was obtained from Phase 1, an enveloping ellipse was defined and convergent iterations begun from zero initial data. The number of iterations required for convergence to a 10^{-4} tolerance are given in Table 3.

In Table 3, the entries marked with a J are the cases where the ordinary Jacobi iteration is convergent (see Table 1) and the spectrum enveloping technique is comparable in efficiency. As discussed later, further improvements in computational efficiency are possible in such cases. In all other cases, the Jacobi iterations diverge and the new technique is quite effective. We note also that the number of iterations required for convergence in the case of $N = 8$ (Table 3) are much smaller than those in Table 2. This is to be expected as the spectrum enveloping ellipses in Table 3 are obtained

TABLE 3
Number of iterations for convergence, spectrum enveloping technique: Problem 1.

	N			
R	8	16	24	32
50	29	36J	62J	106J
100	54	53	51	71J
200	124	343	67	95
500	240	154	90	84
1000	916	302	355	151
2000	905	838	419	277
5000	2934	1586	1009	712

using the algorithm outlined in § 5 whereas the enveloping ellipses in Table 2 were defined in an ad hoc manner.

The second test problem represents convection dominated flows:

$$-\varepsilon(u_{xx}+u_{yy})+u_x=0, \tag{21}$$

in [0, 1] × [0, 1] with boundary conditions

$$\begin{aligned} &u(0,y)=\sin \pi y, \qquad u(1,y)=2\sin \pi y, \qquad 0\leqq x\leqq 1,\\ &u(x,0)=u(x,1)=0, \qquad 0\leqq y\leqq 1. \end{aligned} \tag{22}$$

This problem has been considered by Gartland [3] and Gupta et al. [6]. It is of great physical interest when ε is small. The differential equation (21) is equivalent to (1) with $R=-1/\varepsilon$. The exact solution of (21), (22) is given by

$$u(x,y)=[2e^{(x-1)/2\varepsilon}\sinh \sigma x+e^{x/2\varepsilon}\sinh \sigma(1-x)]\sin \pi y/\sinh \sigma \tag{23}$$

where $\sigma^2=\pi^2+0.25/\varepsilon^2$.

This problem was solved using a uniform $(N+1)\times(N+1)$ mesh for $\varepsilon=0.1, 0.01, 0.005$ and 0.001. In Table 4, we give the number of iterations required by the Jacobi method to either converge to 10^{-4} or diverge to 10^6. An asterisk indicates that the Jacobi iterations diverged.

In Table 5, we give the number of iterations needed for convergence when the algorithm described in § 5 is used to solve the linear system corresponding to (21), (22). It is noted that the spectrum enveloping scheme converges for all values of ε and N. When the standard Jacobi iterations are divergent (entries marked by * in

TABLE 4
Number of iterations for convergence or divergence, Jacobi iterations: Problem 2

	ε			
N	.1	.01	.005	.001
8	50	15*	9*	5*
16	173	35*	15*	6*
24	348	258*	22*	7*
32	564	95	34*	8*
64	1716	230	164	13*

(* = divergence)

TABLE 5
Number of iterations for convergence, spectrum enveloping technique: Problem 2.

	ε			
N	.1	.01	.005	.001
8	52	100	221	1116
16	191	71	138	579
24	370	91	96	395
32	593	95	106	329
64	1785	253	165	267

Table 4), the new iteration scheme converges fairly rapidly as expected. When the Jacobi iterations are convergent (e.g., $\varepsilon = 0.1$; $\varepsilon = 0.01$ with $N = 32$ or 64; $\varepsilon = 0.005$ with $N = 64$), the new iteration scheme converges in about the same or slightly larger number of iterations.

The spectrum enveloping ellipses in all cases in Table 5 used imaginary semi-axis M of length $\geqq 1$. In many of these cases, such as for $\varepsilon = 0.1$, the whole spectrum lies on or near the real line. The dominant eigenvalue as computed using the algorithm outlined in § 5 for $\varepsilon = 0.1$, $N = 8$ is $0.9022 \pm 0.2658i$. With real semi-axis $m = 0.9511$, one needs an enveloping ellipse with imaginary semi-axis $M \geqq 0.8398$. Using $M = 0.8398$, we needed only 44 iterations for convergence which is a saving of 8 iterations. Savings are much more pronounced for larger values of N. The eigenvalue spectra for $\varepsilon = 0.1$, $N \geqq 16$ all lie on the real axis and any value of M (>0) would be satisfactory. As seen in Table 6, using $M = 0.5$ the spectrum enveloping scheme requires approximately 35 percent fewer iterations than using $M = 1$.

Further improvements in the rate of convergence could possibly be obtained by further reducing the value of M. However, this case ($\varepsilon = 0.1$) may not be of great interest since most other iteration methods (Gauss-Seidel, SOR) would probably converge much faster than the Jacobi method.

In the cases of practical importance, we usually find the eigenvalue spectrum to lie away from the real line. In such cases, the conventional iterations often fail to converge and the spectrum enveloping technique could be very effective.

7. Gauss–Seidel method. When the Gauss-Seidel method is used to solve the system of linear equations (3), the eigenvalue spectrum of the iteration matrix is no longer symmetric about the imaginary axis. In many cases the spectrum of the Gauss-Seidel matrix lies outside the infinite strip S. Table 7 contains some of the dominant eigenvalues for Problem 1. These eigenvalues are obtained using Phase 1 of the algorithm given in § 5. In only one case, viz. $N = 16$, $R = 50$, the spectrum enveloping technique of this paper is applicable and the convergence is obtained in 26 iterations.

TABLE 6
Spectrum enveloping data for $\varepsilon = 0.1$: *Problem* 2.

N	Dominant eigenvalue			Number of iterations	
	ξ	η	m	$M = 1$	$M = .5$
16	.8100	0	.8600	191	127
24	.8633	0	9133	370	235
32	.8844	0	.9344	593	372
64	.8902	0	.9402	1785	1159

TABLE 7
Dominant eigenvalues of Gauss-Seidel matrix.
Problem 1.

N	R	ξ	η
8	50	−1.681	0.927
	100	−8.192	1.472
	200	−38.702	3.807
16	50	−0.866	0.262
	100	−2.195	0.960
	200	−11.313	1.390

In the case of Problem 2, the eigenvalue spectrum of the Gauss-Seidel iteration matrix is real for $\varepsilon = 0.1$ and any value of M would be satisfactory. Table 8 contains the dominant eigenvalues of such iteration matrices. Also included is the number of iterations needed for convergence using $M = 1$ and $M = 0.5$. The number of iterations needed by straight Gauss-Seidel method are also given for comparison. It is noted that smaller values of the imaginary semi-axis M helps accelerate the convergence in this case too.

The eigenvalue spectrum for smaller values of ε fails to lie inside S unless N is sufficiently large. The data for $\varepsilon = 0.01$ is given in Table 9. In this case, with $N = 24$ the spectrum lies inside S and the acceleration method of this paper converges in 282 iterations although the standard Gauss-Seidel iterations are divergent. With $N = 32$ the Gauss-Seidel method converges in 28 iterations and spectrum enveloping technique converges in 25.

8. Conclusions. We have presented a method of iteratively obtaining the numerical solution of the convection-diffusion equation (1) for the cases when many of the conventional iteration methods are divergent. The spectrum enveloping technique presented in this paper requires that the real part of the eigenvalues of the basic iteration matrix lie in the interval $(-1, 1)$. When this condition is met, a spectrum enveloping ellipse can be defined and convergent iterations obtained.

It is easy to prove that the eigenvalues of the Jacobi iteration matrix lie in the infinite strip $S = \{|\mathrm{Re}\ \zeta| < 1\}$ when the Reynolds number R is a large constant. We have computationally verified that the Jacobi eigenvalues satisfy this condition even when R is variable. In fact, this is true even when R takes randomly generated large values at each mesh point.

When the spectrum of the basic iteration matrix lies outside the infinite strip S, though to the left of $\mathrm{Re}\,(\zeta) = 1$, a scalar multiplication of the matrix equation can be used to transform this problem into another whose coefficient matrix has its eigenvalue spectrum lying within S. The spectrum enveloping algorithm described in this paper

TABLE 8
Convergence data for Gauss-Seidel iterations: *Problem* 2 ($\varepsilon = 0.1$).

N	$\lvert\xi\rvert$	η	Number of iterations $M = 1$	 $M = 0.5$	 G-S
8	.8514	0	30	31	25
16	.8144	0	110	29	89
24	.7503	0	239	104	184

TABLE 9
Convergence data for Gauss-Seidel matrix: *Problem* 2 ($\varepsilon = 0.01$).

				Number of iterations	
N	ξ	η	M	Spectrum enveloping	G-S
8	−8.203	1.467	*	*	Divergent
16	−2.033	0.930	*	*	Divergent
24	−0.716	0.878	2.4713	282	Divergent
32	0.136	0.741	0.7606	25	28

* Spectrum enveloping not applicable.

can be applied to the transformed problem. However, one must develop an algorithm to compute the parameters of such a transformation. This will be the subject of our future research.

The algorithm described in this paper is robust and has been applied to many test problems. If the basic iterations are convergent, the spectrum enveloping method can converge faster. This method is especially effective when the basic iterations fail to converge.

The Jacobi iteration method is increasingly becoming popular because of its easy adaptability to parallel processors. The spectrum enveloping algorithm is well suited for parallel computation and can be used whether or not the original computational scheme is convergent.

As for the accuracy of the computations, the spectrum enveloping algorithm presented in this paper yields accurate numerical solutions of the algebraic system (3). The numerical solutions exhibit typical oscillations when the convection is dominant, i.e., when the coefficient R in (1) takes large values. This is the well-known property of the central difference approximations [4], [5]. Alternative finite difference approximations of the convection-diffusion equation (1) have recently been proposed by Gupta et al. [6]. These approximations have the property that they yield highly accurate solutions (truncation error of order h^4) and are stable (the eigenvalue spectrum of the iteration matrix is real, lies in $(-1, 1)$, and conventional iteration methods are applicable).

It is well known that direct solvers can be used to solve the systems of algebraic equations when the conventional iteration schemes are not convergent [5]. However, the direct solvers require large memory and possibly large amount of computer time. The spectrum enveloping technique presented in this paper can be used to circumvent some of these difficulties.

REFERENCES

[1] J. DE PILLIS, *How to embrace your spectrum for faster iterative results*, Linear Algebra and Appl., 34 (1980), pp. 125-143.

[2] G. E. FORSYTHE AND W. WASOW, *Finite Difference Methods for Partial Differential Equations*, John Wiley, New York, 1967.

[3] E. C. GARTLAND, *Discrete weighted mean approximation of a model convection diffusion equation*, SIAM J. Sci. Stat. Comput., 3 (1982), pp. 460-472.

[4] P. M. GRESHO AND R. L. LEE, *Don't suppress the wiggles-they're telling you something*, Computers and Fluids, 9 (1981), pp. 223-253.

[5] M. M. GUPTA AND R. MANOHAR, *On the use of central difference scheme for the Navier-Stokes equations*, Internat. J. Numer. Meth. Engrg., 15 (1980), pp. 557-573.

[6] M. M. GUPTA, R. MANOHAR AND J. W. STEPHENSON, *A single cell high order difference scheme for the convection diffusion equation with variable coefficients*, Internat. J. Numer. Meth. Fluids, 4 (1984), pp. 641-651.

[7] L. A. HAGEMAN AND D. M. YOUNG, *Applied Iterative Methods*, Academic Press, New York, 1981.

[8] T. MANTEUFFEL, *The Tchebyshev iteration for nonsymmetric linear systems*, Numer. Math., 28 (1977), pp. 307-327.

[9] W. NIETHAMMER, J. DE PILLIS AND R. S. VARGA, *Convergence of block iterative methods applied to least-squares problems*, Linear Algebra and Appl., 58 (1984), pp. 327-341.

[10] W. NIETHAMMER AND R. S. VARGA, *The analysis of K-step iterative methods for linear systems from summability theory*, Numer. Math., 41 (1983), pp. 177-206.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 4, October 1986

003

THE NULL SPACE PROBLEM I. COMPLEXITY*

THOMAS F. COLEMAN† AND ALEX POTHEN‡

Abstract. The Null Space Problem (NSP) is the following: Given a $t \times n$ matrix A with $t < n$, find a sparsest basis for its null space (a *null basis*). We show that columns in a sparsest null basis correspond to minimal dependent sets of columns of A. Sparsest null bases are characterized by a greedy algorithm that augments a partial basis by a sparsest null vector. Despite this result, (NSP) is NP-hard since finding a sparsest null vector of A is NP-complete. We prove that the related problem of finding a sparsest null basis with an embedded identity matrix is NP-hard too. Finally, we study the zero-nonzero structure of sparsest null bases.

Key words. null basis, null space, sparse matrix, bipartite graph, matching, matroids, conformal decomposition

AMS(MOS) subject classifications. 05, 15, 49, 65, 68

1. Introduction and overview. The development of practical algorithms for the Linear Equality Problem (LEP) is a fundamental concern in numerical optimization. (LEP) can be expressed as

$$\text{minimize}\, f(x)$$
$$\text{subject to}\, Ax = b.$$

Here $f(x)$ is a nonlinear "objective" function $f: \mathbb{R}^n \to \mathbb{R}$, and we assume that f is twice-continuously differentiable. The matrix A has t rows and n columns, $t < n$, and rank $(A) = r$.

Efficient algorithms to solve this problem are needed for two reasons: First, (LEP)s result from mathematical models of several practical optimization problems. Second, (LEP)s occur as subproblems of more general optimization problems. Nonlinearly constrained optimization problems are often solved by linearizing the constraints and solving a succession of resulting (LEP)s. Thus the generalized gradient method, the augmented Lagrangian method, and the projected Lagrangian method to solve these problems are based on efficient algorithms to solve (LEP)s.

One strategy for solving (LEP), the *null space method*, involves two phases: In phase 1, a "feasible" vector y is determined that satisfies $Ay = b$. In phase 2, y is corrected by a vector z in the null space of A that decreases the value of f; that is, $Az = 0$, and $f(y+z) < f(y)$. We set $y := y + z$, and repeat phase 2 until f is small enough in value, or no further reduction in its value can be made.

The correction z can often be chosen so that the algorithm converges at a quadratic rate to a stationary point of f. Let N be a basis for the $(n-r)$-dimensional null space of A (a null basis), $g(y) \in \mathbb{R}^n$ the gradient of f at y, and $H(y) \in \mathbb{R}^{n \times n}$ the Hessian matrix of f at y. We model f about the point y by a quadratic function, and choose $y + z$ to be the minimizer of this model function. This results in the system of equations

$$N^T H(y) N p = -N^T g(y),$$

* Received by the editors July 30, 1984, and in final form November 1, 1985. 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, and a Cornell fellowship to the second author. This work was presented at the SIAM National Meeting at Denver, Colorado June 1983, and at the SIAM Conference on Numerical Optimization at Boulder, Colorado in June 1984.

† Department of Computer Science, Cornell University, Ithaca, New York 14853.

† Center for Applied Mathematics, Cornell University, Ithaca, New York 14853. Present addresc Department of Computer Science, The Pennsylvania State University, University Park, Pennsylvania 16802.

which is solved for the vector p, and then the correction z is computed from the equation $z = Np$.

The system of equations may be solved by computing a factorization of the projected Hessian N^THN when $n-r$ is small. For problems where $n-r$ is large, an iterative technique such as the conjugate gradient method may be used. This is a simplified discussion which ignores several practical issues; Gill, Murray, and Wright (1981) contains a more detailed discussion of (LEP).

Our concern will be with large-scale (LEP). In such problems, the constraint matrix A has a large number of rows and columns. Fortunately, however, most of the matrix elements of A are usually zeros and do not need to be stored. This redeeming feature results from each equation being involved with only a few variables, and each variable occurring only in a small number of equations. Only nonzero elements are stored, allowing large matrices to be processed without exceeding storage capacities of computers. Such matrices, whose zero-nonzero structure can be used to advantage, are *sparse*. Coleman (1984) discusses the various issues that arise in large sparse numerical optimization.

Sparsity in A is good, but is not enough. The null space algorithm needs a representation of a null basis N of A. Such a basis, being a set of $n-r$ vectors that span the null space of A, is not unique, and care needs to be taken to make it as sparse as possible.

With the above discussion to motivate us, we study the Sparse Null Space Basis Problem:

(NSP) A $t \times n$ matrix A with $t < n$ and rank r is given. Find a matrix N with the fewest nonzeros, whose columns span the null space of A.

Hereafter we will abbreviate this to the Null Space Problem. Such an $n \times (n-r)$ matrix N is a *sparsest null basis.*

This paper has four additional sections. We characterize sparsest null bases in § 2 by means of conformal decompositions and matroid theory. The computational complexity of (NSP) and some variants are discussed in § 3. The zero-nonzero structure of sparsest null bases is studied in § 4. In the last section we summarize our results, discuss related work by other researchers, and indicate future research directions. We adopt the notational convention that a term is in *italic* font when it is being defined.

In a second paper, Coleman and Pothen (1985), we will describe our algorithms for computing sparse null bases. These algorithms have two phases: in the first combinatorial phase, a maximum matching in the bipartite graph of A is used to identify the nonzero elements in the null basis. In the second numeric phase, systems of equations are solved to compute numerical values of the nonzeros in the basis. This two-phase strategy makes it possible to efficiently compute sparse null bases. Our computational experience with these algorithms will also be included.

2. A characterization of sparsest null bases. In this section we characterize sparsest null bases by means of a "greedy" algorithm which chooses, at each step, a sparsest possible null vector to be in the basis.

An important concept in what follows is that of a circuit. A linearly dependent set of columns of the matrix A will be called a *dependent set.* A null vector of the matrix A can be obtained from the coefficients of the linear combination. A *circuit* C is a minimal dependent set—i.e., C is dependent, but all proper subsets of C are linearly independent. We will call the null vector associated with a minimal dependent set also a circuit.

ALGORITHM 2.1 (*Greedy Algorithm*). Given a $t \times n$ matrix A with $\operatorname{rank}(A) = r$, find a sparsest null basis N.

```
$N := \emptyset$
for $i = 1, \dots, n - r \rightarrow$
  find a sparsest null vector $n_i$
  such that rank $(n_1, \dots, n_i) = i$.
  $N := N \cup n_i$ rof.
```

THEOREM 2.1 (Optimality Theorem). *The matrix N is a sparsest null basis of A if and only if it can be constructed by the greedy algorithm.*

Algorithm 2.1 is greedy, since it augments the partial null basis at each step by a sparsest null vector linearly independent of those previously chosen. To us Theorem 2.1 is a surprising result; locally greedy strategies seldom lead to globally optimal solutions to optimization problems. We now develop the results needed for its proof.

Let the jth component of a vector x be denoted by $(x)_j$. (This should not be confused with the notation for a vector, say n_i.) We define the *support* of x, $S(x)$, to be

$$S(x) = \{j: (x)_j \neq 0\}.$$

By definition, if c is a circuit, there cannot exist a null vector x with $S(x) \subset S(c)$.

LEMMA 2.2. If c, d are circuits of A, and $S(c) = S(d)$, then c is a scalar multiple of d.

Proof. Suppose the lemma is false. Then we can pick a scalar λ such that $(c)_i - \lambda(d)_i = 0$, for some $i \in S(c)$. But then $S(c - \lambda d) \subset S(c)$, and c is not minimal. □

Hence circuits of A are unique to within a multiplicative constant. We now introduce a linear algebraic concept from network flow theory, conformal decomposition, studied first by Camion (1968), Fulkerson (1968), and Rockefellar (1969). Lemmas 2.2 through 2.4 follow immediately from their work.

A vector x *conforms* to a vector y if

$$(x)_j \neq 0 \Rightarrow ((y)_j \neq 0, \text{ and } \operatorname{sgn}\{(x)_j\} = \operatorname{sgn}\{(y)_j\})$$

where sgn denotes the sign function. For example, let

$$\operatorname{sgn}(x) = (+0-0+0), \qquad \operatorname{sgn}(y) = (++-0+-),$$

then x conforms to y, but y does not conform to x. Note that if x conforms to y, then $S(x) \subseteq S(y)$.

LEMMA 2.3. *Given a null vector n, there exists a circuit c that conforms to it.*

Proof. Again, the proof is by contradiction. Choose a null vector x with the smallest $|S(x)|$ such that no circuit of A conforms to it. Let c be a circuit with $S(c) \subset S(x)$. Define the set

$$J = \{j: (c)_j \neq 0, \text{and } (c)_j \text{ and } (x)_j \text{ disagree in sign}\}.$$

J is not the empty set, else c would conform to x. Let

$$a = \min_{j \in J} -\frac{(x)_j}{(c)_j}.$$

Consider the vector $z = x + ac$. By construction, z conforms to x, and $S(z) \subset S(x)$. By the selection of x there is a circuit d that conforms to z. But then d conforms to x. □

We can now apply Lemma 2.3 repeatedly to get

LEMMA 2.4. *A null vector x can be expanded in a sum of distinct circuits*

$$x = c_1 + \cdots + c_p,$$

where each circuit c_i conforms to x.

The above expansion is the *conformal decomposition* of a null vector of A; it is not necessarily unique. A more general decomposition exists for a vector of any subspace of $\mathbb{R}^n$, and is discussed by Camion, Fulkerson and Rockefellar. We can now use Lemma 2.4 to prove that we need concern ourselves only with circuits to solve (NSP).

THEOREM 2.5. *Each sparsest null vector n_i chosen by the greedy algorithm is a circuit.*

Proof. The proof is by induction on i. The result is clearly true for n_1. By the inductive hypothesis, assume that the theorem is true for all n_j, where $1 \leq j < i$.

Suppose that n_i is not a circuit. Conformally decompose n_i into a sum of circuits. At least one of the circuits in this sum, say c, must be linearly independent of $(n_1, \cdots, n_{i-1})$ since n_i is independent of them. Since n_i is not a circuit, $S(c) \subset S(n_i)$, and c is a sparser null vector than n_i which the algorithm could have chosen at this step. □

A similar argument can be used to prove

THEOREM 2.6. *Each column of a sparsest null basis N is a circuit.*

Theorem 2.6 states that the only dependent sets of interest in (NSP) are circuits. Since the greedy algorithm chooses only circuits by Theorem 2.5, the possibility now looms that it could find a sparsest null basis. As Theorem 2.1 states, this suspicion is correct; and a stronger result holds, namely, every sparsest null basis can be found by the greedy algorithm.

We now introduce the matroid concepts used to prove Theorem 2.1. Let E be a finite set. Some of the subsets of E are defined to be *independent*; a subset of E that is not independent is *dependent*. Let

$$H = \{I \subseteq E \colon I \text{ is independent}\}.$$

We consider the situation when the independent sets satisfy the following two properties:

(M1) All subsets of an independent set are independent. (The empty set is independent by this property if H is not empty.)

(M2)- Let I_p and I_{p+1} be independent sets with p and $p+1$ elements respectively. Then there is an element $e \in I_{p+1} \backslash I_p$ such that $I_p + e$ is independent.

Let the family of independent sets H satisfy (M1) and (M2). Then the tuple $M = (C, H)$ is defined to be a *matroid* (Welsh (1976)).

The reader may find it convenient to think of E as the set of columns of a matrix. An independent subset of E has linearly independent columns. By linear algebra, one can establish that both (M1) and (M2) hold. Hence M is a matroid, and we call it the matroid generated by the columns of the matrix.

A minimal dependent set of a matroid is called a *circuit*. Thus far we have used the word circuit to denote a minimal linearly dependent set of columns of a matrix. This usage is consistent with the definition of a circuit of a matroid. What we call a circuit of a matrix is indeed a circuit of the matroid generated by the columns of the matrix.

A *maximal independent set* is an independent set all supersets of which are dependent. We call such a set a *basis* of M. Every basis of M has the same size, which is called its *rank*.

Proof of Theorem 2.1. By Theorems 2.5 and 2.6 we can restrict our attention to circuits of A. Since A has n columns, it has only a finite number of circuits. Let C be the circuit matrix whose columns are all the circuits of A. Thus

$$C=(c_1,\cdots,c_q).$$

Let M be the matroid generated by the columns of C. To each circuit c_i, assign the positive integer weight $|S(c_i)|$. Algorithm 2.1 is equivalent to choosing a basis of minimum weight for the circuit matroid M. Theorem 2.1 now follows from two well-known results on matroids:

(1) The matroid greedy algorithm constructs a basis of minimum weight.

(2) The weight of the kth smallest element of such a basis is no bigger than the kth smallest element of any other independent set (Lawler (1976)). □

Unfortunately, the proof of Theorem 2.1 does not lead immediately to a polynomial time algorithm to solve (NSP). The difficulty is that a matrix A of n columns and t rows might have $O(n^t)$ circuits.

3. The complexity of (NSP) and its variants. In the previous section, we showed that a sparsest null basis can be constructed by a greedy algorithm. Hence we consider the following strategy to solve (NSP): design a polynomial time algorithm for one step of the greedy algorithm. This latter algorithm would choose a sparsest circuit linearly independent of circuits chosen in previous steps. If we could design such an algorithm, then $n-r$ applications of it to the matrix A will solve (NSP).

Unfortunately, such a happy prospect is unlikely; we now discuss the reason why. The greedy algorithm chooses a circuit of minimum cardinality in its first step. We call such a circuit a *minimum circuit.* Theorem 3.1 states that the minimum circuit problem is NP-complete. Hence it is as hard as any of the problems in the class NP. For the reader unfamiliar with this terrain, Garey and Johnson (1979) is an excellent introduction to the theory of NP-completeness. Theorems 3.1 and 3.2 were proved independently by L. J. Stockmeyer, and his proofs may be found in McCormick (1983).

THEOREM 3.1 (Minimum Circuit Theorem). *Given a positive integer k, it is* NP-*complete to find a circuit of A of cardinality k or less.*

We omit our proof since our reduction is similar to Stockmeyer's. Theorem 3.1 leads to an easy proof that (NSP) is NP-hard. We do not know if (NSP) is in NP.

THEOREM 3.2 (Sparsest Null Basis Theorem). *Given a positive integer k, it is* NP-*hard to find a null basis of A with k or fewer nonzeros.*

Proof. By Theorem 2.1, every sparsest null basis contains a minimum circuit. By Theorem 3.1, it is NP-complete to find a minimum circuit. □

If A is restricted to be the vertex-edge incidence matrix of a graph $G=(V,E)$, a minimum circuit can be found in $O(|V|\,|E|)$ time by an algorithm of Itai and Rodeh (1978). In this situation, a minimum circuit corresponds to a cycle in the graph with the minimum number of edges. Matroids generated by vertex-edge incidence matrices of graphs are called *graphic matroids.*

Every matroid has a dual defined on the same ground set C. A basis of the dual matroid is the complement of a basis of the primal matroid. A matroid dual to a graphic matroid is *cographic.* Minimum circuits of cographic matroids correspond to minimum cuts in the graph; these can also be found in polynomial time.

A matrix A is *totally unimodular* if every subdeterminant of A is either $+1$, -1, or 0. The matroid generated by such a matrix is called a *totally unimodular matroid.* Seymour (1980) has shown that any totally unimodular matroid can be decomposed by a polynomial time algorithm into a matroid sum of graphic matroids, cographic matroids, and copies of a special matroid on ten elements. It follows that minimum circuits of totally unimodular matroids can be determined in polynomial time.

We have shown that constructing a null basis of a matrix A with the maximum number of nonzeros is also NP-hard when the columns in the basis are circuits of A.

THEOREM 3.3. *Given a positive integer k, it is* NP-*hard to find null basis of A with k or more nonzeros, if each column in the basis is a circuit.*

The proof is by the restriction of A to vertex-edge incidence matrices of graphs, and uses the result that finding a basis with the maximum number of edges for the cycle space of a graph is NP-complete. A proof is presented in Pothen (1984).

Since (NSP) is NP-hard, we cannot expect to construct sparsest null bases by a polynomial time algorithm. Hence we lower our sights in terms of sparsity, and ask how hard it is to construct a sparsest null basis with a prescribed zero-nonzero structure.

Current null space algorithms for (LEP) use the variable-reduction technique proposed by Wolfe (1962) to construct null bases. Let A_r denote any r linearly independent rows of A. The matrix A_r is partitioned (after possible column permutations) as

$$A_r = (M \quad U),$$

where M is a $r \times r$ nonsingular matrix. Then we construct the matrix

$$N = \begin{pmatrix} -M^{-1}U \\ I_{n-r} \end{pmatrix},$$

where I_{n-r} is the identity matrix of dimension $n-r$. Since $AN=0$, the columns of N are null vectors of A. Each of the last $n-r$ rows of N has only one nonzero in it, and so linear combinations of the columns of N cannot produce the zero vector. Hence N is a null basis. We call a basis with an embedded identity submatrix a *fundamental null basis.*

We formally state the Fundamental Null Space Problem:

(FNSP) Given a $t \times n$ matrix A of rank r and a positive integer k, find a fundamental null basis N with k or fewer nonzeros.

THEOREM 3.4. (FNSP) *is* NP-*hard.*

The proof of this theorem uses a result on spanning trees of graphs. We now develop the concepts needed for the proof.

Let $G=(V, E)$ be a connected graph on ν vertices and ε edges with vertex-edge incidence matrix $M(G)$. A *cycle* in G is a sequence of distinct vertices $v_1, \ldots, v_{k-1}, v_k \equiv v_1$, where $(v_{i-1}, v_i) \in E$ for $i = 2, \ldots, k$. Denote the edge incidence vector of a cycle by Γ, with component γ_i equal to 1 if e_i is an edge of the cycle, and 0 otherwise. Since each vertex in the cycle is an endpoint of exactly two edges, we have

$$M(G)\Gamma^T = 0,$$

over the binary field GF(2). Thus Γ is a null vector of $M(G)$. Further, since the omission of any edge in the cycle will violate the equation, Γ is a circuit of $M(G)$. Since every circuit of M has zero or two edges incident on each vertex of G, there is a one-to-one correspondence between a cycle of G and a circuit of $M(G)$ over GF (2).

However, our interest is with circuits over the real field. But the restriction to arithmetic over GF(2) is easily removed. Let D be a directed graph obtained by arbitrarily directing the edges of G. The vertex-edge incidence matrix of D, $M(D)$ has in the column of the directed edge $\{u, v\}$ the entry $+1$ in the row of v, -1 in the row of u, and 0 in all other rows. A *cycle* in D is defined to be a cycle in G with an arbitrary orientation. Let $\Gamma(D)$ be the edge incidence vector of a cycle in D, with component γ_i equal to $+1$ if e_i is an edge in the cycle and the orientations of the cycle and e_i

agree, -1 if e_i is an edge in the cycle and the orientations disagree, and 0 if e_i is not an edge in the cycle. We have

$$M(D)\Gamma(D)^T=0,$$

where the arithmetic is now over the real field. Further, there is a one-to-one correspondence between a cycle of D and a circuit of $M(D)$.

We now extend this correspondence to one between a fundamental null basis of $M(D)$ and an appropriate graph concept. A *spanning tree* T of an undirected graph G is a connected subgraph with ν vertices and $\nu-1$ edges. Each nontree edge e creates a unique cycle $C(T, e)$ in the subgraph $T+e$. We call $C(T, e)$ the *fundamental cycle* created by e with respect to T. Since T has $\nu-1$ edges, there are $\omega(G)\equiv\varepsilon-\nu+1$ nontree edges. Hence there are $\omega(G)$ fundamental cycles with respect to T. *The fundamental cycle matrix* $\Phi(G)$ has $\omega(G)$ rows and ε columns, with element ϕ_{ij} equal to 1 if e_j is an edge of the cycle Φ_i, and 0 otherwise. If the edges of T are numbered from 1 to $\nu-1$, and the nontree edges from ν to ε, then $\Phi(G)$ has the structure $\Phi(G)=(\Phi_{11}\ I)$.

Let D be a directed graph obtained from G as before. A *spanning tree* of D is defined to be a spanning tree of G. The *fundamental cycle matrix of D*, $\Phi(D)$, has element ϕ_{ij} equal to $+1$ if e_j is an edge of Φ_i and their orientations agree, -1 if e_j is an edge of Φ_i and their orientations disagree, and 0 if e_j is not an edge of Φ_i. Thus for any spanning tree T, $\Phi(G)$ and $\Phi(D)$ have the same structure. Since each row of $\Phi(D)$ corresponds to a cycle in D, we have

$$M(D)\Phi(D)^T=0,$$

where the arithmetic is over the real field. Hence $\Phi(D)^T$ is a fundamental null basis for $M(D)$.

Proof of Theorem 3.4. Restrict A to vertex-edge incidence matrices of directed graphs. A sparsest fundamental null basis of A now corresponds to a sparsest fundamental cycle matrix of the associated directed graph. The latter is equivalent to finding a sparsest fundamental cycle matrix of the undirected graph obtained by ignoring the directions of the edges. This last problem is that of finding a spanning-tree that minimizes the total number of edges in the set of fundamental cycles with respect to it. This problem is NP-complete; proofs may be found in Deo, Prabhu, and Krishnamoorthy (1982) and Pothen (1984). □

A fundamental basis for the row space of A has the structure $(I_r\ B)$ and corresponds to a fundamental null basis

$$\begin{pmatrix}-B\\ I_{n-r}\end{pmatrix}$$

with only a constant change in the number of nonzeros. Hence we have

COROLLARY 3.5. *Given a positive integer k, it is* NP-*hard to find a fundamental row space basis of A with k or fewer nonzeros.*

In contrast, finding a (nonfundamental) sparsest row space basis can be done in polynomial time (Hoffman and McCormick (1984)) when the matrix A satisfies a nondegeneracy assumption called the matching property.

4. The structure of sparsest null bases. Any algorithm for constructing a null basis has to ensure that the set of $n-r$ null vectors chosen is linearly independent. Constructing a fundamental null basis makes this easy to do. However, sparsest null bases need not be fundamental. We may be constrained to construct relatively dense fundamental bases where sparse nonfundamental null bases may exist.

But what zero-nonzero structure (hereafter *structure*) should a sparsest null basis have? By Theorem 4.2 below, a set of m vectors is linearly independent for all nonzero values of its nonzero elements if and only if it has an embedded upper triangular submatrix of dimension m. In what follows, let V be a matrix with n rows and m columns, with $n > m$. Distinguish some elements of V as nonzeros and the rest as zeros. By a value of a matrix we mean an assignment of nonzero numerical values to its nonzero elements.

LEMMA 4.1. *If V has at least two nonzeros in each row, then there exists a nonzero vector x, and a value for V, such that $Vx = 0$.*

Proof. Let row i have $|r_i| \geqq 2$ nonzeros. We assign to any $|r_i|-1$ nonzeros the value$+1$, and to the remaining element the value $1-|r_i|$. We do this for all the rows of V, and choose $x = (1 \dots 1)^T$. □

THEOREM 4.2. *V has rank m for all values if and only if it can be permuted to the following structure:*

$$V = \begin{pmatrix} B \\ U_m \end{pmatrix},$$

where U_m is an $m \times m$ upper triangular matrix with nonzero diagonal elements.

Proof. The if part is obvious. We prove the only if part. Suppose that V has rank m, but does not have the structure claimed. Permute the rows and columns of V so that it has the structure

$$V = \begin{pmatrix} B & C \\ O & R \end{pmatrix},$$

where R is upper triangular and maximal with respect to this property. Since R is maximal, B has at least two nonzeros in each of its rows. By Lemma 4.1, we can now find a vector x and numeric values for the nonzeros of B so that $Bx = 0$. Since

$$V\begin{pmatrix} x \\ 0 \end{pmatrix} = 0,$$

V does not have rank m. This contradiction proves the theorem. □

It may appear from this theorem that a sparsest null basis should have an embedded upper triangular matrix. This would be true if we could assign any value to N. But, we are not free to do so. We can assign any value to A; then, once the structure of N is chosen, the values of the columns of N are uniquely determined to within a multiplicative constant.

Theorem 4.3 concerns the structure of a sparsest null basis. This result is a matroid generalization of a theorem on cycles in graphs proved by Stepanets (1964). We shall denote the set of columns of the matrix A also by A. Let $n(a_j)$ be a circuit of minimum cardinality containing the column a_j. The reader may find Fig. 1 helpful to follow the proof of this theorem.

THEOREM 4.3 (Generalized Stepanets Theorem). *Let the columns $a_1, \cdots, a_k$ be chosen such that*

$$a_1 \in A,$$
$$a_2 \in A \backslash n(a_1), \quad \dots,$$
$$a_k \in A \Big\backslash \bigcup_{j=1}^{k-1} n(a_j).$$

There exists a sparsest null basis N among whose columns are the circuits $n(a_1), \cdots, n(a_k)$.

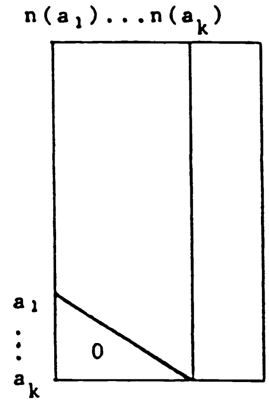


FIG. 1. *The sparsest null basis N.*

Proof. We prove the theorem by induction on $n(a_i)$. Let $q=n-r$, and denote the set $(P\backslash p)\cup n$ by $P-p+n$.

Let $P=(p_1\cdots p_q)$ be a sparsest null basis of A. Since P is a basis, we can expand $n(a_1)$ as

$$n(a_1)=c_1p_1+\cdots+c_mp_m.$$

We assume that all the coefficients in this expansion are nonzero. Of the circuits in this equation, there must exist at least one circuit, say p_h, which contains a_1. Consider the system

$$P_1=P-p_h+n(a_1).$$

Clearly P_1 is a null basis. Further, since $n(a_1)$ has minimum cardinality over circuits containing a_1, P_1 is a sparsest null basis.

For the inductive step, assume that P_{j-1} is a sparsest null basis of A, having among its columns $n(a_1),\ldots,n(a_{j-1})$, where each a_i is chosen as claimed. We choose $n(a_j)$ to be a circuit of minimum cardinality containing a_j. Expand $n(a_j)$ in the basis P_{j-1},

$$n(a_j)=c_1p_1+\cdots+c_mp_m,$$

where again each of the coefficients is nonzero. There is at least one circuit in this equation, say p_h, which contains a_j. The circuit p_h cannot be any one of $n(a_1),\cdots,n(a_{j-1})$ by the choice of a_j. Consider now the system

$$P_j=P_{j-1}-p_h+n(a_j).$$

As before, P_j is a sparsest null basis of A.

We take N to be P_k. This completes the proof. □

There is some $k\leq n-r$ for which choosing a column a_{k+1} is not possible, since the first k circuits contain all columns of A.

COROLLARY 4.4. *If $k=n-r$ in Theorem 4.3, then the system of circuits $n(a_1),\cdots,n(a_k)$ is a sparsest null basis N of A.*

In this case, N has an upper triangular submatrix with $n-r$ columns. We call such a basis a *triangular null basis.* Thus, some sparsest null bases are triangular.

5. Conclusions. We have formulated the Null Space Problem, and the Fundamental Null Space Problem. We have shown that only circuits can be columns in a sparsest null basis, and that such a basis can be characterized by a matroid greedy algorithm. However, (NSP) is NP-hard since a sparsest null basis contains a minimum circuit and finding a minimum circuit is NP-complete. Constructing a sparsest fundamental null basis is also an NP-hard problem. Hence the use of approximation algorithms to solve (NSP) is justified.

A fundamental null basis ensures linear independence of the set of null vectors chosen. We have extended this observation to show that a set of vectors is linearly independent for all values if and only if it has an embedded upper triangular submatrix with nonzeros on the diagonal. This can be used in approximation algorithms to construct triangular null bases for which linear independence of the null vectors is again easy to ensure.

A problem related to (NSP) is that of finding a set of cycles with the fewest edges that spans the cycle space of a graph. Note that our proof technique for the NP-hardness of (NSP) does not extend to this problem, since cycles with the fewest edges can be found in polynomial time. The complexity of this problem is open (Johnson (1985)). However, the problem of finding a set of fundamental cycles with the fewest edges that span the cycle space of a graph is NP-complete (Deo, Prabhu and Krishnamoorthy (1982), Pothen (1984)). The problem of finding a set of cycles with maximum number of edges spanning the cycle space of a graph is NP-complete (Pothen (1984)).

In Coleman and Pothen (1985), we will show how circuits can be constructed from a maximum matching in the bipartite graph of the matrix A. This algorithm can be used repeatedly to construct fundamental null bases. Here the sparsity of the basis turns out to depend only on the partition of the columns of A into the matched and unmatched sets. Various heuristic strategies for finding particular matchings are used to obtain sparse null bases.

By varying the matching while constructing null vectors, a triangular null basis can be obtained. Such bases can be potentially sparser than fundamental null bases; however, this increase in sparsity is achieved at greater computational cost.

We briefly mention recent work related to (NSP). Berry, Heath, Kaneko, Lawo, Plemmons and Ward (1985) have implemented a refined version of a "turnback algorithm", proposed initially by Topcu (1979), that constructs sparse null bases for large sparse, banded A. This algorithm uses an initial numeric factorization of A to identify subsets of columns that could become dependent sets in the $n-r$ null vectors. In a second turnback phase, a numeric factorization on each dependent set is performed to obtain circuits. Their numerical results on several problems arising from finite element models in structural engineering show that they obtain null bases with the same degree of sparsity as the input matrices. Berry and Plemmons (1985) have implemented a parallel version of this algorithm on a Denelcor HEP computer. Gilbert and Heath (1986) have implemented several algorithms for computing sparse null bases; some of these are closer in spirit to the ones we have designed. For instance, in one of their algorithms, they construct a triangular null basis; the columns in each circuit are identified by matching methods.

Much work remains to be done. An important numerical consideration is the condition number of the null basis. To this end, algorithms that can compromise some degree of sparsity for better conditioned null bases will need to be developed. Other sparsity criteria than the one used in this paper need to be studied. We mention one such in closing. An *implicit null basis* is a representation for the null basis as a product of a sequence of elementary matrices (e.g., Givens rotations), with the sequence of elementary matrices being stored. A sparse implicit null basis has relatively few elementary matrices in the sequence. One direction in which we plan to continue this research is in developing sparse implicit orthogonal null bases.

Acknowledgments. We wish to thank Professors John Gilbert and Bob Bland for several valuable discussions and Ms. Helen Johnston for technical assistance.

REFERENCES

M. W. BERRY, M. T. HEATH, I. KANEKO, M. LAWO, R. J. PLEMMONS AND R. C. WARD, *An Algorithm to compute a sparse basis of the null space*, Numer. Math., 47 (1985), pp. 483-504.

M. W. BERRY AND R. J. PLEMMONS, *Computing a banded basis of the null space on the Denelcor* HEP *multiprocessor*, in Proc. AMS/SIAM Summer Conference on Linear Algebra in Systems Theory, AMS Series on Contemp. Math., 1985.

P. CAMION, *Modules unimodulaires*, J. Combin. Theory, 4 (1968), pp. 301-362.

THOMAS F. COLEMAN, *Large Sparse Numerical Optimization*, Lecture Notes in Computer Science 165, Springer-Verlag, Berlin, 1984.

THOMAS F. COLEMAN AND ALEX POTHEN, *The sparse null space basis problem* II. *Algorithms*, (in preparation), 1985.

NARSINGH DEO, G. M. PRABHU AND M. S. KRISHNAMOORTHY, *Algorithms for generating fundamental cycles in a graph*, ACM Trans. Math. Software, 8 (1982), pp. 26-42.

D. R. FULKERSON, *Networks, frames, blocking systems*, in Mathematics of the Decision Sciences, Vol. 2, G. B. Dantzig and A. F. Veinott, eds, American Mathematical Society, Providence, RI, 1968, pp. 303-334.

M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability*: *A Guide to the Theory of* NP-*Completeness*, W. H. Freeman, San Francisco, 1979.

JOHN R. GILBERT AND MICHAEL T. HEATH, *Computing a sparse basis for the null space*, Cornell University and Oak Ridge National Laboratory Technical Report, 1986.

PHILIP E. GILL, WALTER MURRAY AND MARGARET H. WRIGHT, *Practical Optimization*, Academic Press, New York, 1981.

ALAN J. HOFFMAN AND S. THOMAS MCCORMICK, *A fast algorithm for making matrices optimally sparse*, in Progress in Combinatorial Optimization, W. R. Pulleyblank, ed., Academic Press, New York, 1984.

ALON ITAI AND MICHAEL RODEH, *Finding a minimum circuit in a graph*, SIAM J. Comput., 7 (1978), pp. 413-423.

DAVID S. JOHNSON, *The* NP-*Completeness column*: *An ongoing guide*, J. Algorithms, 6 (1985), pp. 145-159.

EUGENE L. LAWLER, *Combinatorial Optimization*: *Networks and Matroids*, Holt, Rinehart, and Winston, New York, 1976.

S. THOMAS MCCORMICK, *A combinatorial approach to some sparse matrix problems*, SOL 83-5, Ph.D. Thesis, Stanford Univ., Stanford, CA, 1983.

ALEX POTHEN, *Sparse null bases and marriage theorems*, Ph.D. Thesis, Cornell Univ. Ithaca, NY, 1984.

R. T. ROCKEFELLAR, *The elementary vectors of subspaces of* $\mathbb{R}^n$, in Combinatorial Mathematics and its Applications, R. C. Bose and T. A. Dowling, eds., Univ. North Carolina Press, Chapel Hill, NC 1969, pp. 104-127.

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

G. F. STEPANETS, *Basis systems of vector cycles with extremal properties in graphs*, Upsekhi Mat. Nauk., 19 (1964), pp. 2, 171-175. (In Russian.)

A. TOPCU, *A contribution to the systematic analysis of finite element structures through the force method*, Ph.D. Thesis, Univ. of Essen, Essen, Germany, 1979. (In German.)

D. J. A. WELSH, *Matroid Theory*, Academic Press, London, 1976.

PHILIP WOLFE, *The reduced gradient method*, The RAND Corporation, 1962, unpublished manuscript.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 4, October 1986

COLUMN LU FACTORIZATION WITH PIVOTING ON A MESSAGE-PASSING MULTIPROCESSOR*

GEORGE J. DAVIS†

Abstract. A column-oriented algorithm is presented for LU factorization with partial pivoting. Two different mappings of columns to processors are considered. Forward and backsubstitution algorithms to use the factorization for solving linear systems are developed. Timing data, including processor utilization and load balancing, are provided by a hypercube simulator.

Key words. LU factorization, message-passing multiprocessors

AMS(MOS) subject classifications. 15A23, 6SFO5

1. Overview. In the past several years there has been considerable activity aimed at producing and utilizing various types of multiprocessing architectures. Many such computers already exist, and others are in their final development stages. This activity has presented new challenges in the area of algorithm development. Problems of task partitioning, scheduling, memory access and synchronization must be addressed. The overall goal is to provide an implementation which best exploits its computational environment. Balancing the workload among processors to maximize their individual activity is one of the best ways to reduce overall execution time.

Examining the details of a particular architecture can give valuable insight on how a program should be designed. Strategies for balancing the workload on a shared-memory multiprocessor may be very different from those in a distributed memory environment. The processor interconnection network also plays a part in deciding how to partition a given problem. In this paper we consider a message-passing environment in general, and the hypercube connection network in particular for two important reasons. First, the hypercube is among the first multiprocessors to become commercially available and thus many people will be learning about parallel algorithms while using hypercubes. Second, many other of the common interconnection strategies can be embedded in a hypercube network. Thus algorithms designed for these other topologies can also be run on the hypercube.

This paper presents a column oriented algorithm for LU factorization with partial pivoting for use on a message-passing multiprocessor. A description of the hypercube multiprocessor and its simulator follow in § 2. Two different mappings of columns to multiprocessors are provided, and these are described (along with the algorithms themselves) in § 3. Details of the forward and backsubstitutions are provided in § 4. The final section contains a timing analysis of the test examples, as provided by the simulator.

2. The hypercube multiprocessor. One of the more popular message-passing multiprocessors currently available is the hypercube (or cosmic cube [6]). It consists of $p=2^N$ independent processors, or *nodes*. N is called the *dimension* of the hypercube.

* Received by the editors August 29, 1985 and in revised form December 10, 1985. This work was supported by the Applied Mathematical Sciences Research Program, Department of Energy Office of Energy Research and was performed while the author was a summer visitor to the Mathematical Sciences Section, Engineering Physics and Mathematics Division of the Oak Ridge National Laboratory, which is operated by Martin Marietta Energy Systems, Inc., for the U.S. Department of Energy under contract DE-AC05-84OR21400.

† Mathematics and Computer Science, Georgia State University, Atlanta, Georgia 30303-3083.

Each node is a sequential computer possessing its own local memory and its own operating system. An important feature of this operating system is its ability to send, receive and route messages which pass through the node. Nodes function concurrently and independently of each other, making the hypercube a multiple-instruction, multiple-data (MIMD) machine. Machines with $p=64$ now exist, and those with $p=1024$ are feasible in the near future.

In addition to the p nodes there is also one additional processor called the *host.* The job of the host is to spawn processes on the nodes, to collect information from them and to handle input/output. We assume (as in the Intel iPSC hypercube) that there is a two-way communication link from the host to each node. This link is not used for node-to-node message-passing. In other types of hypercubes, the host may be connected to only one or a few of the nodes.

The nodes are identified with the binary representation of the numbers 0 through $p-1$. Communication between nodes is done by passing messages rather than by sharing memory. Individual nodes may communicate only with their immediate neighbors. Two individual nodes are immediate neighbors if their binary identifiers differ by exactly one bit. For example, when $p=8$, node number 010 is directly connected to node number 110 but not to 111. If these binary tags are thought of as ordered triples corresponding to coordinates of a cube in three-dimensional space, nodes are directly connected if and only if there is an edge between their corresponding coordinates. A message to be sent from 010 to 111 must pass through other intermediate nodes. Details about how messages can be routed between nodes can be found in [4].

We assume the existence of two message-passing primitives *send* and *await.* The *send* is an asynchronous command which sends a message of a designated length to a designated destination and returns control to the sending process. The *await* causes a process to suspend execution until a message of a designated type arrives. Messages arriving at a destination before the execution of an appropriate *await* are queued. Two other primitives which will be of use are: *cubedim*( ), which returns the dimension of the hypercube, and *mynode*( ) which allows a node to determine its binary identifier. All four of these functions are very similar to what is available on actual hypercubes.

The codes contained in this paper were developed and run on a hypercube simulator which runs on a VAX 11/780 under the Unix operating system at Oak Ridge National Laboratory. The simulator was written by T. H. Dunigan of Oak Ridge [2], and is based on the multitasking kernel of E. D. Brooks [1]. Individual processes are modeled within the framework of the VAX in such a way that they have access only to their local memory. Data are sent from process to process by passing messages. The simulator provides a comprehensive trace file containing a log of all messages sent and received as well as the wait states of the processes. A clock time measured in VAX instructions tags each event. Communication delays can be effectively modeled by adjusting the message arrival times. The trace file can be piped to other programs which analyze processor utilization and provide performance graphs.

The simulator can be invoked in one of two modes. In the first mode, after the host has spawned processes on the nodes, node 0 executes until it encounters an *await.* Then node 1 awakens and executes until it encounters an *await.* Control is passed from node to node sequentially until node $p-1$ yields control back to node 0. This mode, although not a realistic model of the hypercube, can be quite useful for uncovering logic and synchronization errors. Print generated from these runs is also quite useful, although the trace file with the clock is not available.

The second mode more closely resembles the actual operation of the cube, with the appearance of all processes running asynchronously. Print statements for debugging purposes need to be carefully selected, as it is quite common for output from different processes to become interleaved. Subtle errors in timing more often are encountered at this point. The importance of the trace file cannot be overestimated. In the actual cube a program could stall due to processes with unsatisfied *awaits*, yet the user would have no idea which process was waiting for what. By recording the wait states of the nodes, the trace file is a great help in resolving these kinds of problems.

The simulator allows the possibility of spawning different processes on different nodes, although our node programs are identical. This tends to make the codes slightly longer, due to special-case startups and shutdowns, but the host process is greatly simplified. All codes were written in C.

3. The algorithms. Let A be the nonsingular coefficient matrix of a square linear system $Ax = b$, whose order is n. We explore the idea of sending a subset of the columns of A to each processor. Obviously the matrix can be partitioned in other ways, for example, by rows or by blocks. A study of partitioning by rows for a general matrix can be found in [3]. Several strategies for partitioning for the Cholesky factorization if A is symmetric can be found in [5], with more specific information about column partitions in [4]. Discussions on the relative merits of each approach can be found in all these references.

Having decided to map columns to processors, the next decision involves the nature of the mapping. There are three natural ways to send columns to the nodes: blocking, wrapping and reflecting. To simplify the discussion, we assume that the number of processors (p) divides the order of the matrix (n), although this is not a requirement of the algorithm.

Blocking requires that the first (n/p) columns are sent to node 0, the next (n/p) to node 1, and so on. The obvious disadvantage of such an approach lies in the fact that after node 0 has factored its columns and sent its information to the other nodes, it remains idle for the rest of the factorization. A block mapping will keep the last processor busy through most of the computation, but will not fully use the earlier ones. We will only consider the other two partitioning schemes.

3.1. Wrap mapping. In the wrap mapping of columns to processors, column k is sent to processor $k \pmod p$. Note that since the p processors are labeled 0 through $p-1$, we also choose to label the columns of the matrix 0 through $n-1$. Although this is a bit of a nuisance for most FORTRAN programmers, it fits in quite naturally with C conventions.

Having the columns wrapped on the processors, we now need to describe a message-passing strategy which will both assure proper synchronization and maximum processor utilization. By processor utilization we always mean the relative amount of the time that a node is active, as opposed to awaiting messages. The schematic given in Fig. 1 will be used to describe this algorithm.

We have chosen a "send to next" procedure as opposed to a general broadcast designed for the hypercube. Throughout this paper, the "next node" is taken to be that node which needs a given piece of information first. For example, the multipliers from the elimination of column i are immediately needed by that node possessing column $i+1$. The reason for not broadcasting is that depending on the size of the messages and the overhead in a *send*, it was quite possible (if not inevitable) that pivots and multipliers would arrive at nodes out of order. Although this can be remedied by tagging each message with a column number, or giving each column number a different message type, for clearer understanding of the algorithm we have chosen our

```
Host process:
    Generate (or load) $n$, $A$ and $b$
    Spawn processes on the nodes.
    Send appropriate columns to nodes.
    Await each pivot and column of $L$ and $U$ from nodes.
        Apply pivots and $L$ to $b$.
        Form transpose of $U$.
    Send appropriate rows of $U$ and modified $b$ to nodes.
    Await each component of the finished solution $x$.

Node process:
    Await columns of $A$ from host.
    If node 0:
        Determine first pivot index and first column of $L$.
    While factorization not done:
        Await a previous pivot index and set of multipliers.
        Forward these messages to the next node.
        Apply row interchanges to all columns.
        If this node has the next column to be processed:
            Apply multipliers to active column only.
            Select new pivot, and send to next node.
            Form new column of $L$.
            Send multipliers to next node.
            Apply received and generated multipliers to remaining columns.
        Else:
            Apply received multipliers to remaining columns.

    Await rows of $U$ and solution of $Ly = b$ from host.
    If this node has the last row of $U$:
        Compute last component of $x$ and send to next node.
    While backsubstitution not done:
        Await a previous component of $x$.
        Forward to the next node.
        If this node has the next row in the backsubstitution:
            Determine next component of $x$.
            Send component to next node.
        Update all remaining components of $x$.
```

FIG. 1. *Wrap mapping.*

method. It turns out that by using this strategy, we are effectively treating the hypercube connection network as a ring: node k communicates with node $k+1$ (mode p).

Several other remarks are in order. First it should be noted that the right-hand side b resides on the host, and the forward substitution is done at the same time the factorization is being completed. After each column of A is pivoted and factored into L and U, the factored column is sent to the host. Right-hand side b is interchanged and the multipliers in L are applied to it.

While the host is waiting for a new pivot and factored column to arrive, it also transposes the newly received column of U. The reason for this is that backsubstitution cannot be performed in parallel with U stored on the nodes by columns.

As an example, assume an 8×8 matrix A, with columns numbered 0 through 7, has been wrapped onto 4 processors and the factorization has been completed in place. Suppose that the solution to $Ly = b$ has been done on the host. The logical place to send y is to processor 3, as it has the $(7, 7)$ element of U. Component 7 of the solution can now be computed, and an updated y can be sent out. The only processor which can now become active is number 2, as all the others require information not yet available. Component 6 of the solution and a further updated y can now be sent, again with only one other processor able to become active.

The backsubstitution is thus sequential. Indeed, experiments with doing the factorization and forward substitution of a single b during the factorization, and the backsubstitution sequentially revealed that the backsubstitution can easily amount to 30% of the total execution time when $n > 100$. With U transposed, the parallel backsubstitution is less than 4% of the total execution time for the same sized problems.

It should be stressed that the algorithm in its current form solves $Ax = b$ with the right-hand side(s) processed sequentially during the factorization. If A were to be factored and used at different times with different right-hand sides, the advantage of the concurrent forward substitution would be lost. Also, if several right-hand sides were to be processed from the outset, the work of the host may not be completely masked by the factorization.

As an alternative, a possible implementation could transpose both L and U and send their appropriate rows back to the nodes, so that the forward substitution would also be parallel and follow the same logic as the backsubstitution. We would expect that the two solution phases would consume about 8% of the total time with problem size as above.

Finally, note that the forward substitution does not require elaborate synchronization, although in this scheme pivots and multipliers will arrive in sequential order.

With regard to the factorization itself, a reading of Fig. 1 reveals that the parallel codes for the node processes possess a logic quite different from serial codes. An individual node must realize which columns it has, which ones have already been processed, and when it possesses the next column to be factored. A node declares itself to be done if it has processed its last column.

Several features were designed into the node process code to maximize processor utilization. For example, suppose a node possesses column number k and has just received the multipliers from column $k-1$. Rather than apply these mutipliers to all of its subsequent columns, the node applies them only to column k while it is selecting the new pivot index. The pivot index is immediately sent to the next node. Then the new multipliers are calculated and immediately sent on. Only then will the node apply the multipliers from column $k-1$, the pivot from column k and the multipliers from column k to the remaining columns it possesses. This strategy allows the generated information to be sent out as soon as it becomes available. The same kind of logic is found in the backsubstitution.

3.2. Reflection mapping. In the reflection mapping of columns to processors, columns 0 through $p-1$ are sent to processors 0 through $p-1$, just as in wrapping. The next p columns, however, are sent to the processors in reverse order from $p-1$ to 0. The direction of mapping then reverses for each subsequent p columns. For example, if $n = 20$ and $p = 4$:

processor 0 has columns: 0, 7, 8, 15, 16
processor 1 has columns: 1, 6, 9, 14, 17
processor 2 has columns: 2, 5, 10, 13, 18
processor 3 has columns: 3, 4, 11, 12, 19.

The general formula can be described as follows:

if int (j/p) is even: send column j to processor $j \pmod p$
if int (j/p) is odd: send column j to processor $(p-1)-(j \pmod p)$.

We have presented a schematic for the reflected algorithm in Fig. 2.

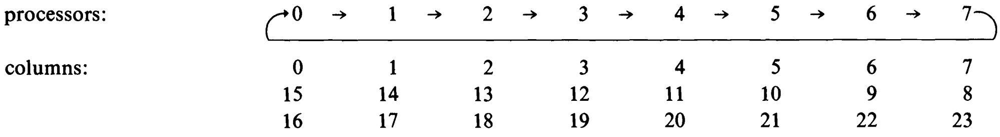
```
Node process:
    *Await* columns of $A$ from host.
    *If* node 0:
        Determine first pivot index and first column of $L$.
    *While* factorization not done:
        *Await* a previous pivot index and set of multipliers.
        Forward these messages to the next node.
        Apply row interchanges to all columns.
        *If* this node has the next column to be processed:
            Apply multipliers to active column only.
            Select new pivot, and send to next node.
            Form new column of $L$.
            *Send* multipliers to next node.
            Apply received and generated multipliers to remaining columns.
            *If* this node has the next column to be processed:
                Apply multipliers to active column only.
                Select new pivot, and send to next node.
                Form new column of $L$.
                *Send* multipliers to next node.
                Apply received and two generated sets of multipliers to the
                  remaining columns.
            *Else*:
                Apply received and generated multipliers to remaining columns.
        *Else*:
            Apply received multipliers to remaining columns.

    *Await* rows of $U$ and solution of $Ly = b$ from host.
    *If* this node has the last row of $U$:
        Compute last component of $x$ and send to next node.
    *While* backsubstitution not done:
        *Await* a previous component of $x$.
        Forward to the next node.
        *If* this node has the next row in the backsubstitution:
            Determine next component of $x$.
            *Send* component to next node.
            *If* this node has the next row in the backsubstitution:
                Determine next component of $x$.
                *Send* component to next node.
        Update all remaining components of $x$.
```

FIG. 2. *Reflection mapping.*

The host process is the same as in Fig. 1, and is not shown. One can see that the overall structure is much the same, but that there is one additional inner loop. The reason for this is that processors 0 and $p-1$ possess columns which must be factored in immediate succession. With an eye toward maximum utilization, we again send out information as soon as it becomes available, without waiting for all columns in a given node to be modified.

The synchronization strategy for the reflection mapping is more complex. We will consider at length an example where $p = 8$ and n is at least 24, which can adequately illustrate the method.

Using the right-type communication shown in Fig. 3, serious difficulties are encountered. Consider first the propagation of the pivot index and multiplier information for columns 7 and 8, which reside on node 7. This information must eventually be sent to all other processors. We illustrate the distribution of the columns in Fig. 3, with the arrows indicating the routing of messages from node to node.



processors:	0 →	1 →	2 →	3 →	4 →	5 →	6 →	7
columns:	0	1	2	3	4	5	6	7
	15	14	13	12	11	10	9	8
	16	17	18	19	20	21	22	23

FIG. 3. *Ring-type communication.*

We see that the information from node 7 is first needed on node 6, for the elimination of column 9. If the messages are routed from node 7 to node 0, node 6 will be in an *await* state until the needed information cycles through all the other nodes. This can result in a significant decrease in overall processor utilization.

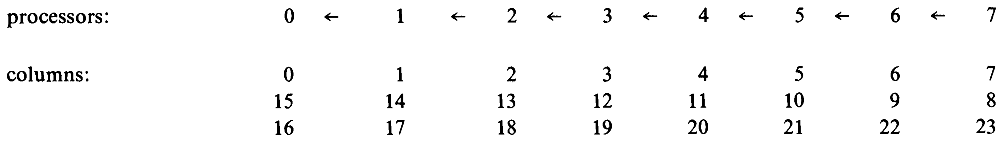
A better communication strategy would have pivots and multipliers sent where they are needed next. This implies that in the above example, node 7 sends its information to node 6, where it is subsequently forwarded to nodes 5, 4, etc. This is shown in Fig. 4.

A final illustration helps to solidify the strategy. Consider the propagation of information from column 12, which resides on node 3. The next nodes to require this information are, in order, 2, 1 and 0. Now the next processor to need the pivots and multipliers from column 12 is node 4, followed by 5, 6 and 7. This is the routing which is employed (see Fig. 5).

In this strategy, the information is always sent where it is needed next. Although this routing seems fairly complicated at first, once the formula for the destination is worked out, it does not involve much coding. Routing for the backsubstitution follows similar logic.

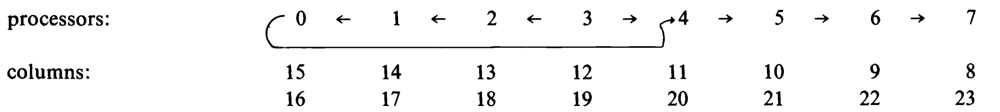
It cannot be overstated that synchronization of a column-oriented algorithm is a critical problem. It is very often the case that subtle timing errors will not be evidenced on small-order examples which can be checked by hand. This scheme, while having a tendency to slow down the propagation of information though the nodes, assures proper and correct execution.

4. Forward and backsubstitution. It has been noted that the forward substitution (solution of $Ly = b$) is done on the host processor. Preliminary experiments on some hypercubes indicate that this may not be the best strategy to reduce overall execution time. Node-to-host communication may be significantly slower than node-to-node communication. If this is the case, then the forward substitution could become a significant fraction of the total work.



processors:	0 ←	1 ←	2 ←	3 ←	4 ←	5 ←	6 ←	7
columns:	0	1	2	3	4	5	6	7
	15	14	13	12	11	10	9	8
	16	17	18	19	20	21	22	23

FIG. 4. *Reflection communication: Propagation of multipliers from node* 7.



processors:	0 ←	1 ←	2 ←	3 →	4 →	5 →	6 →	7
columns:	15	14	13	12	11	10	9	8
	16	17	18	19	20	21	22	23

FIG. 5. *Reflection communication: Propagation of multipliers from node* 3, *column* 12.

The work of transposing and re-sending the upper triangular factor U may also not be masked by the factorization. However, algorithms are currently being developed for efficiently transposing a matrix entirely on the nodes. These techniques are particularly important when the matrix U is so large that it cannot entirely be accumulated on the host. In the setting where U is more modestly sized and the node-to-host and node-to-node communication times are roughly equivalent, results more like our simulator runs can be expected.

We have assumed that the host processor is essentially equivalent in computational power to the nodes. Although this may not be the case in an actual hypercube, the volume of work expected of the host in these algorithms is not significant. Besides sending and receiving messages from the nodes, the host is expected only to modify the right-hand side(s) b.

The message-passing strategies for the backsubstitution phase follows the logic in Figs. 3-5, but with the arrows of communications reversed. The key idea is to send the newly acquired components of solution x where they will be needed next.

5. Analysis of the test examples. The codes were tested on a number of linear systems with varying orders, using varying numbers of processors. Matrices filled with random numbers were generated and multiplied by a given solution vector x to create problems $Ax = b$ with known solutions. These A and b were passed to the factorization and solution algorithms on the simulator at Oak Ridge. Trace files for each run were recorded. All of the matrices generated were nonsingular and well conditioned, and the algorithms gave accurate solutions.

Tables 1 and 2 show the sizes of some of the test examples with statistics generated from examining the trace files. For each problem, a time measured in the number of VAX operations is given, along with percentages of node and host utilization. A node utilization percentage of 46% means that the node was in a busy state for 46% of the time, and in an *await* state for the remaining 54%. A busy state involves computation, receiving or sending messages.

TABLE 1

Wrap mapping.

n	p	time (vax ops)	node util range (%)	host util (%)	n	p	time (vax ops)	node util range (%)	host util (%)
8	4	3689	46-64	53	64	16	120497	52-71	57
8	2	3918	73-80	48	64	8	180111	70-81	38
16	8	9774	38-60	59	64	4	305003	78-82	22
16	4	11379	60-72	50	64	2	424741	95-95	16
16	2	14056	83-86	40	128	64	369655	24-50	71
32	16	29230	20-57	73	128	32	440405	50-71	59
32	8	35218	55-71	54	128	16	672442	70-82	39
32	4	50840	71-79	37	128	8	1157869	80-87	22
32	2	68944	90-92	27	128	4	2143351	81-84	12
64	32	97104	28-54	71	128	2	2980744	97-98	9

A number of interesting facts are revealed by these tables. First, the reflection mapping had a tendency to balance the overall percentages of node utilization better than the wrap mapping. Focus, for example, on the $n = 128$, $p = 32$ case. Reflection

TABLE 2

Reflection mapping.

n	p	time (vax ops)	node util range (%)	host util (%)
8	4	3656	52-58	57
8	2	3695	75-76	55
16	8	9602	46-53	63
16	4	11011	66-69	54
16	2	13541	82-86	44
32	16	28498	41-50	70
32	8	34464	62-66	57
32	4	48601	77-78	40
32	2	68653	88-92	28
64	32	94622	36-48	74
64	16	118685	60-64	59
64	8	175916	76-77	40
64	4	29639	81-83	24
64	2	427302	93-95	16
128	64	343364	33-46	77
128	32	435453	58-64	60
128	16	663479	76-78	40
128	8	1139338	84-85	23
128	4	2108581	83-84	12
128	2	2998556	96-98	9

mapping produced node utilizations from 58% to 64%, while wrap mapping produced a range of 50% to 71%. The average node utilizations (in this case 62% vs 61%) were always very close, with a slight edge to reflection. This is the same experience reported in [3].

As one would expect, the utilization increases as the ratio n/p increases, with percentages as high as 98% when $n = 128$ and $p = 2$. Each processor has more data and therefore more work to do once it receives a pivot index and a set of multipliers. On the other hand, the host utilization drops, as the forward substitution becomes a smaller fraction of the total time.

Tables 3 through 6 give more detailed statistics for two of the examples, $n = 64$ with $p = 8$, and $n = 128$ with $p = 8$. There was nothing special about these two cases other than that they are large enough to show some typical behavior. Under the heading "tid" is given the number of each node, with "h" standing for host. The next two columns labelled "start" and "end" give the VAX clock time at which the node process began and ended. Note that the host does not start at time 0. This is due to the fact that the matrix and right-hand side are being generated before the factorization begins.

The "duration" column gives the total amount of time (measured in VAX instructions) that a given node was active, and the "busy" column records the amount of time in a non-*await* state. "Busy" divided by "duration" therefore gives the nodal utilization. Note that the total number of sends does not match the total number of receives. This is due to the fact that as the factorization and backsolve are completed,

TABLE 3

Wrap mapping: $n = 64$, $p = 8$. *Nodal utilization* 75% *Nodal + host utilization* 71% *sends* 1788 *recvs* 1764.

tid	start	end	duration	busy	utiliz	sends	recvs
h	103258	283369	180111	68515	38%	144	192
0	103651	283345	179694	125640	70%	202	193
1	103817	283259	179442	125897	70%	203	194
2	103983	283143	179160	129160	72%	204	195
3	104149	283051	178902	132282	74%	205	196
4	104315	282835	178520	135369	76%	206	197
5	104481	282743	178262	138349	78%	207	198
6	104647	282627	177980	141224	79%	208	199
7	104813	282535	177722	143789	81%	209	200

TABLE 4

Wrap mapping: $n = 128, p = 8$. *Nodal utilization* 83% *Nodal + host utilization* 77% *sends* 3644 *recvs* 3620.

tid	start	end	duration	busy	utiliz	sends	recvs
h	411290	1569159	1157869	259636	22%	272	384
0	412067	1569135	1157068	931013	80%	418	401
1	412385	1569049	1156664	930398	80%	419	402
2	412703	1568933	1156230	942954	82%	420	403
3	413021	1568841	1155820	955379	83%	421	404
4	413339	1568625	1155286	967522	84%	422	405
5	413657	1568533	1154876	979516	85%	423	406
6	413975	1568417	1154442	991306	86%	424	407
7	414293	1568325	1154032	1002590	87%	425	408

TABLE 5

Reflection mapping: $n = 64, p = 8$. *Nodal utilization* 77% *Nodal + host utilization* 73% *sends* 1617 *recvs* 1596.

tid	start	end	duration	busy	utiliz	sends	recvs
h	103258	279174	175916	69849	40%	144	192
0	103651	279151	175500	134877	77%	120	186
1	103895	279042	175147	133483	76%	213	183
2	104139	278916	174777	133825	77%	210	180
3	104383	278814	174431	134058	77%	207	177
4	104627	278608	173981	134220	77%	204	174
5	104871	278506	173635	134312	77%	201	171
6	105115	278380	173265	134270	77%	198	168
7	105359	278290	172931	133175	77%	120	165

TABLE 6

Reflection mapping: $n = 128, p = 8$. *Nodal utilization* 85% *Nodal + host utilization* 78% *sends* 3281 *recvs* 3260.

tid	start	end	duration	busy	utiliz	sends	recvs
h	411290	1550628	1139338	262389	23%	272	384
0	412067	1550605	1138538	970227	85%	240	370
1	412551	1550496	1137945	960420	84%	429	367
2	413035	1550370	1137335	961221	85%	426	364
3	413519	1550268	1136749	961863	85%	423	361
4	414003	1550062	1136059	962264	85%	420	358
5	414487	1549960	1135473	962472	85%	417	355
6	414971	1549834	1134863	962522	85%	414	352
7	415455	1549744	1134289	960143	85%	240	349

nodes may be sending information to other nodes which have already shut down. This somewhat inelegant ending could be repaired with several more lines of code to avoid unnecessary *sends.* We again see the better load balance and slightly higher overall utilization of reflection.

Finally we present performance graphs (Figs. 6–9) for the cases detailed in Tables 3 through 6. The numbers on the top refer to the sizes of n and p. Thus "n64p8wrap" refers to $n = 64$, and $p = 8$ with the wrap mapping, and "n64p8ref" refers to the same size problem with reflection mapping.

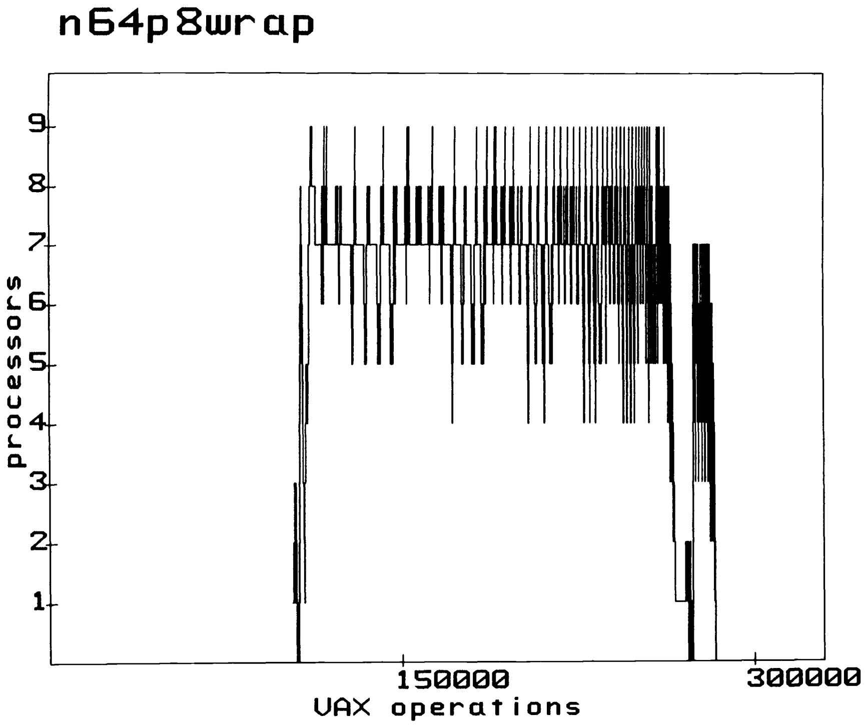

n64p8wrap
processors
1
2
3
4
5
6
7
8
9
150000
300000
VAX operations

FIG. 6

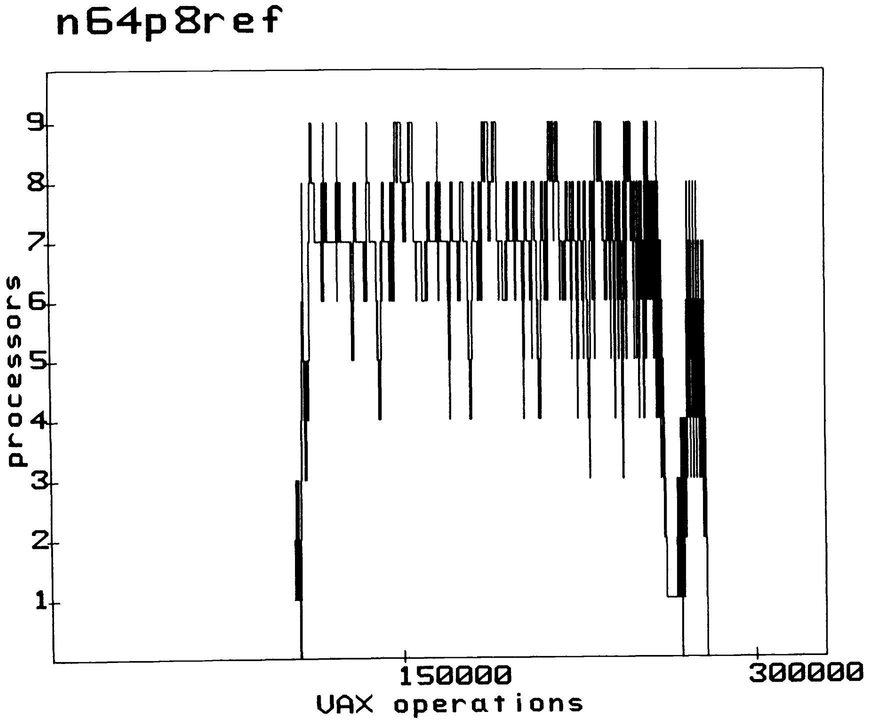

n64p8ref
processors
1
2
3
4
5
6
7
8
9
150000
300000
VAX operations

FIG. 7

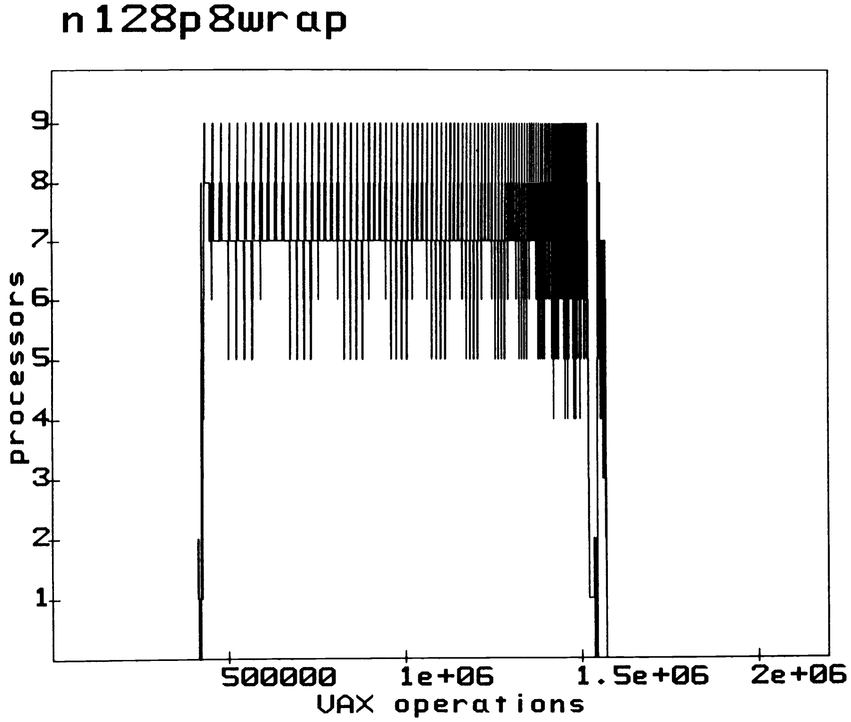

n128p8wrap
processors
1
2
3
4
5
6
7
8
9
500000
1e+06
1.5e+06
2e+06
VAX operations

FIG. 8

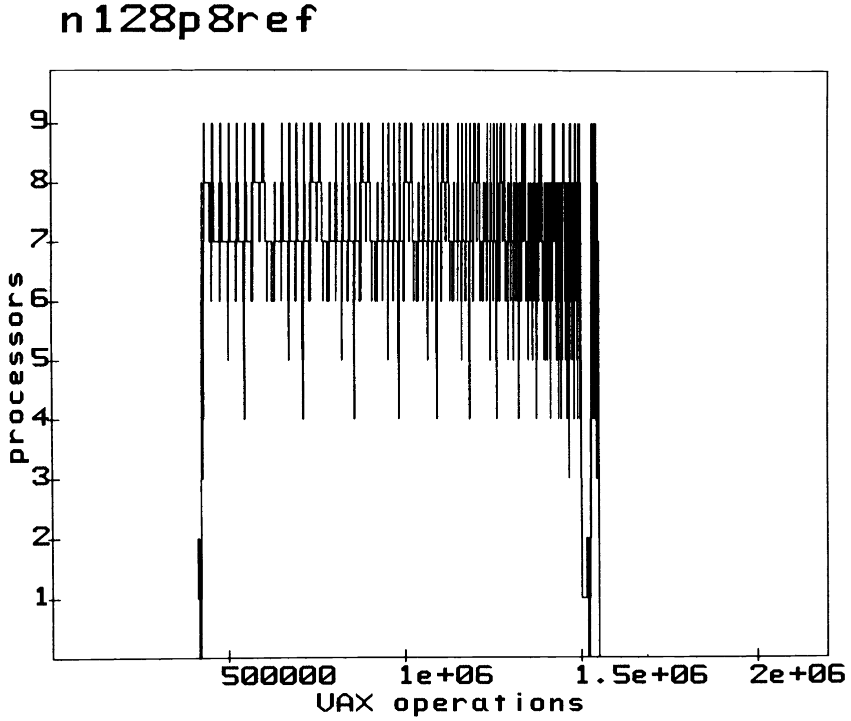

n128p8ref
processors
1
2
3
4
5
6
7
8
9
500000
1e+06
1.5e+06
2e+06
VAX operations

FIG. 9

The graphs show the number of processors which are active at any given time during the algorithm. The initial gap with no processors active corresponds to the generation of the problem. Even though there are only eight nodes, at various times the graphs show nine processors busy, corresponding to the additional activity of the host. The jagged dropoffs toward the end of the graphs indicate the backsubstitution phase, and how little time it takes with respect to the factorization. These kinds of plots, especially for larger problems, given an important qualitative feel for how the algorithms control the activity of the nodes.

It can be seen that these algorithms do a good job of keeping the individual nodes busy. A "perfect" performance plot would jump up quickly to eight (or nine) processes active and remain there until the computations are completed. In these examples, the time during which six or more processes are active comprises a large fraction of the total effort.

Acknowledgments. The author is gratefully indebted to Al Geist, Mike Heath and Bob Funderlic of Oak Ridge National Laboratory for many helpful discussions and suggestions regarding this work.

REFERENCES

[1] E.D. BROOKS, *A multitasking kernel for the* C *and Fortran programming languages*, Lawrence Livermore National Laboratory Technical Report UCID-20167, 1984.

[2] T.H. DUNIGAN, *A parallel processor simulator*, in preparation.

[3] G. A. GEIST, *Efficient parallel LU factorization with pivoting on a hypercube multiprocessor*, Oak Ridge National Laboratory Technical Report ORNL-6211, 1985.

[4] G. A. GEIST AND M. T. HEATH, *Parallel Cholesky factorization on a hypercube multiprocessor*, Oak Ridge National Laboratory Technical Report ORNL-6190, 1985.

[5] M. T. HEATH, *Parallel Cholesky factorization in message-passing multiprocessor environments*, Oak Ridge National Laboratory Technical Report ORNL-6150, 1985.

[6] C. L. SEITZ, *The Cosmic Cube*, Comm. ACM, 28 (1985), pp. 22-33.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 4, October 1986

NEIGHBORHOODS OF DOMINANT CONVERGENCE FOR THE SSOR METHOD*

MICHAEL NEUMANN†

Abstract. Let A be an $n\times n$ nonsingular irreducible 3-cyclic H-matrix and let J^A, L_ω^A and S_ω^A denote, respectively, the Jacobi, the SOR, and the SSOR iteration matrices associated with A. In this paper we show that if the spectral radius $\rho(|J^A|)\in(0, r_0)$, where r_0 is the unique root of the cubic $17r^3+r^2-r-1$ in the interval $(0,1)$, then there exists a neighborhood $\Omega_{\omega(A)}$ of $\omega(A):=2/(1+\rho(|J^A|)$ such that

$$\rho(S_\omega^A)<|\omega-1|\leqq\rho(L_\omega^A)\quad\forall\,\omega\in\Omega_{\omega(A)}.$$

Key words. iterative methods, H-matrices

AMS(MOS) subject classifications. Primary 65F10, secondary 15A06, 15A18

1. Introduction. Let $A=(a_{ij})$ be an $n\times n$ complex nonsingular H-matrix whose diagonal entries are, without loss of generality, all unity and write

$$A=I-L-U, \tag{1.1}$$

where L and U are, respectively, a strictly lower and strictly upper triangular matrices. Two well-known results from the theory of successive overrelaxation (SOR) are that

$$\rho(L_\omega^A)\leqq|1-\omega|+\omega\rho(|J^A|)\quad\forall\omega\in(0,\omega(A)] \tag{1.2}$$

with $\rho(L_\omega^A)<1\ \forall\omega\in(0,\omega(A))$ and that

$$\rho(S_\omega^A)\leqq|1-\omega|+\omega\rho(|J^A|)\quad\forall\omega\in(0,\omega(A)] \tag{1.3}$$

with $\rho(S_\omega^A)<1\ \forall\omega\in(0,\omega(A))$, where

$$L_\omega^A:=(I-\omega L)^{-1}[(1-\omega)I+\omega U] \tag{1.4}$$

is the SOR *iteration matrix associated with* A and

$$S_\omega^A:=(I-\omega U)^{-1}[(1-\omega)I+\omega L](I-\omega L)^{-1}[(1-\omega)I+\omega U] \tag{1.5}$$

is the *symmetric* SOR (SSOR) *iteration matrix associated with* A, respectively, and where

$$\omega(A):=\frac{2}{1+\rho(|J^A|)}, \tag{1.6}$$

J^A being the *Jacobi iteration matrix associated with* A, namely

$$J^A:=L+U. \tag{1.7}$$

Note that, as $\rho(|J^A|)<1$, $\omega(A)>1$.

We mention that (1.2) was shown first to be true for nonsingular M-matrices A by Kahan in [5]. Kahan's results were later shown to hold for the total class of H-matrices by Kulisch, see [6]. The inequality (1.3) for the SSOR method was proven by Alefeld and Varga in [1].

* Received by the editors August 12, 1985, and in revised form January 9, 1986.

† Department of Mathematics, The University of Connecticut, Storrs, Connecticut 06268. The work of the author was supported in part by the National Science Foundation under grant MCS-8400879.

The questions of whether $\rho(L_\omega^A)$ and $\rho(S_\omega^A)$ attain the value 1 at $\omega=\omega(A)$ have been investigated in [8], [9], [10], and [12]. In [12] it was shown that the upper bound (1.2) is sharp for each $\omega\in(0,\omega(A)]$ on the total class of nonsingular H-matrices of all orders. In [9] it was shown that if A is an $n\times n$ p-cyclic irreducible nonsingular M-matrix in the canonical form

$$A=\begin{bmatrix} I & & & B_{1p}\\ B_{21} & \ddots & & \\ & \ddots & \ddots & \\ & & B_{p,p-1} & I\end{bmatrix} \tag{1.8}$$

then $\rho(L_{\omega(A)}^A)<1$; whereas for $A_1=A^T$, $\rho(L_{\omega(A_1)}^{A_1})<1$ if and only if p is even. From the point of view of the SOR theory, the difference in the behavior of these cases results from the fact that A in (1.8) is consistently ordered in the sense of Varga [14, p. 101], while $A_1=A^T$ is inconsistently ordered.

Continuing, in [10] it was shown that

$$\rho(S_{\omega(A)}^A)<1 \quad \text{unless } \rho(|J^A|)=0. \tag{1.9}$$

Furthermore, two additional facts were noted:

(a) The first fact is the following observation:

OBSERVATION 1.1 [10, § 3]. *Suppose that A is an $n\times n$ nonsingular M-matrix. Then*

$$\rho(S_\omega^A)\leqq\rho(L_\omega^A) \quad \forall\omega\in(0,1], \tag{1.10}$$

with strict inequality holding in (1.10) *if A is* (*also*) *irreducible.*

We comment that the proof of this inequality rests on a comparison theorem, due to Woźnicki [16], for the asymptotic convergence rate of two iteration matrices resulting from two regular splittings of a monotone matrix A. (For a more accessible exposition of Woźnicki's results and further generalizations see [4] and [7].)

(b) The second fact which was noted in [7] is that in many examples of H-matrices A,

$$\rho(S_{\omega(A)}^A)<\rho(L_{\omega(A)}^A) \tag{1.11}$$

with a substantial difference at times.

In regards to (1.11) *only heuristic* reasons were advanced in [10] to justify the inequality and one purpose of this paper is to provide an interval $(0, r_0)\in(0,1)$ such that for all nonsingular 3-cyclic irreducible H-matrices A with $\rho(|J^A|\in(0,r_0))$, (1.11) holds. In fact the following is a corollary to our main Theorem 2.2.

COROLLARY 1.2. *Let A be a* 3-*cyclic nonsingular irreducible H-matrix and let r_0 be the unique root in the interval* (0, 1) *of the polynomial*

$$g(r)=17r^3+r^2-r-1. \tag{1.12}$$

If $\rho(|J^A|)\in(0,r_0)$, then there exists a real open neighborhood $\Omega_{\omega(A)}$ of $\omega(A)$ such that

$$\rho(S_\omega^A)<|\omega-1|\leqq\rho(L_\omega^A) \quad \forall\omega\in\Omega_{(A)}. \tag{1.13}$$

In view of the upper bound for the spectral radius of S_ω^A given in (1.13), we shall think of the neighborhood $\Omega_{\omega(A)}$ referred to in Corollary 1.2 as a *neighborhood of dominant convergence of the* SSOR *method over the* SOR *method.*

We mention that the unique root r_0 in (0, 1) of the cubic (1.12) is roughly equal to 0.4181928.

Our main results are developed in § 2. We finally make the following remarks: (i) In a recent survey on preconditioned iterative methods, O. Axelsson, see [2], compares the performance of various iterative schemes as preconditioning strategies for use in conjunction with Chebyshev acceleration techniques or with conjugate gradient methods. He notes, citing various references, the particular effectiveness of the SSOR method as a preconditioning method and its insensitivity to the choice of relaxation parameter ω. (ii) In a recent paper [8], Neumaier and Varga have given the exact domain of convergence, as a function of ω and $r=\rho(|J^A|)$, for the SSOR method for all nonsingular H-matrices.

Remark. The notation and terminology used in this paper are very standard to the literature on nonnegative matrices, M-matrices, H-matrices, diagonal dominance and iterative methods. The interested reader who is not familiar with the subject is referred to the following excellent texts by A. Berman and R. J. Plemmons [3], by R. S. Varga [14] and by D. M. Young [17] for explanation of the notation employed here and for further background material.

2. Dominance neighborhoods for SSOR. Let A be an $n\times n$ nonsingular H-matrix. Our starting point is simple: Since a well-known lower bound on the spectral radius of the SOR iteration matrix due to Kahan [5] is

$$|\omega-1|\leqq\rho(L_\omega^A)\quad\forall\omega\neq 0, \tag{2.1}$$

we raise the following question:

What is the largest interval $(0,r)\subseteq(0,1)$ *in* r *such that*

$$\rho(S_{\omega(A)}^A)<|\omega(A)-1|=\omega(A)-1, \tag{2.2}$$

for all nonsingular H-matrices A such that $\rho(|J^A|)\in(0,r)$?

As a first step towards answering this question we mention that in an earlier version [11] of [10] a method for simplifying the investigation of the SSOR matrix S_ω^A was suggested. On S_ω^A perform the similarity transformation to obtain the matrix $\hat{S}_\omega^A$ as follows:

$$\begin{aligned}\hat{S}_\omega^A&=(I-\omega U)S_\omega^A(I-\omega U)^{-1}\\&=(I-\omega L)^{-1}[(1-\omega)I+\omega L](I-\omega U)^{-1}[(1-\omega)I+\omega U].\end{aligned}\tag{2.3}$$

The advantage of working with (2.3) over S_ω^A is that now $\hat{S}_\omega^A$ is a product of a (block) lower triangular matrix by a (block) upper triangular matrix, thus making the (block) structure of $\hat{S}_\omega^A$ somewhat easier to handle than the (block) structure of S_ω^A.

Next, on applying the Neumann expansion in (2.3), we observe that $\hat{S}_\omega^A$ admits the following representation

$$\hat{S}_\omega^A=(1-\omega)^2I+R_\omega^A. \tag{2.4}$$

Set

$$\tilde{S}_\omega^A:=(1-\omega)^2I+|R_\omega^A|. \tag{2.5}$$

Then $|\hat{S}_\omega^A|\leqq\tilde{S}_\omega^A$ so that by (2.3) and the Perron–Frobenius theory,

$$\rho(S_\omega^A)=\rho(\hat{S}_\omega^A)\leqq\rho(\tilde{S}_\omega^A). \tag{2.6}$$

Thus, in view of (2.2) and (2.6), a cruder question than the one which was posed above is the following:

What is the largest interval $(0,r)\subseteq(0,1)$ *in* r *such that*

$$\rho(\tilde{S}_{\omega(A)}^A)<\omega(A)-1 \tag{2.7}$$

for all nonsingular and irreducible H-matrices A with $\rho(|J^A|)\in(0,r)$?

It is immediate from (2.5) that

$$\rho(\tilde{S}^A_\omega)=(1-\omega)^2+\rho(|R^A_\omega|).$$

Thus, for $\omega=\omega(A)$, inequality (2.7) holds if and only if

$$\rho(|R^A_{\omega(A)}|)<\omega(A)-1-(1-\omega(A))^2. \tag{2.8}$$

Now put $r:=\rho(|J^A|)$ in which case, according to (1.6), $\omega(A)=2/(1+r)$. Then upon substituting $2/(1+r)$ for $\omega(A)$ in (2.8) we see that inequality (2.7) holds if and only if

$$\rho(|R^A_{\omega(A)}|)<\frac{2r(1-r)}{(1+r)^2}=:p(r). \tag{2.9}$$

From here on we shall assume that A is a nonsingular block 3-cyclic irreducible H-matrix. As mentioned in the introduction, whether A is consistently or inconsistently ordered can effect the attainment of the value 1 at $\omega=\omega(A)$ by the spectral radius of the SOR matrix, but that in either case, unless $\rho(|J^A|)=0$, the spectral radius of the SSOR matrix does not attain the value 1 at $\omega=\omega(A)$. We shall suppose that A has the form

$$A=\begin{bmatrix} I & 0 & B_{13} \\ B_{21} & I & 0 \\ 0 & B_{32} & I \end{bmatrix}; \tag{2.10}$$

however all our results are valid for the case $A_1=A^T$ using similar lines of argument. Thus the SSOR method justifies its name to a degree since its behavior is reasonably independent of the block structure of A.

For A given in (2.10) the computation of $\hat{S}^A_\omega$ of (2.4) yields that

(2.11)

$$\hat{S}^A_\omega=(1-\omega)^2I$$
$$+\begin{bmatrix} 0 & 0 & \omega(1-\omega)(\omega-2)B_{13} \\ \omega(1-\omega)(\omega-2)B_{21} & 0 & \omega^2(\omega-2)^2B_{21}B_{13} \\ \omega^2(1-\omega)(2-\omega)B_{32}B_{21} & \omega(1-\omega)(\omega-2)B_{21} & \omega^3(2-\omega)(\omega-2)B_{32}B_{21}B_{13} \end{bmatrix}.$$

Now set

$$\tilde{R}^A_{\omega(A)}:=\begin{bmatrix} 0 & 0 & \omega(A)\delta(A)C_{13} \\ \omega(A)\delta(A)C_{21} & 0 & \omega^2(A)\mu^2(A)C_{21}C_{13} \\ \omega^2(A)\delta(A)C_{32}C_{21} & \omega(A)\delta(A)C_{32} & \omega^3(A)\mu^2(A)C \end{bmatrix}, \tag{2.12}$$

where

$$C_{13}:=|B_{13}|,\quad C_{21}:=|B_{21}|,\quad C_{32}:=|B_{32}|\quad\text{and}\quad C=C_{32}C_{21}C_{13}, \tag{2.13}$$

$$\delta(A):=(\omega(A)-1)(2-\omega(A)), \tag{2.14}$$

and

$$\mu(A):=2-\omega(A). \tag{2.15}$$

Thus by (2.4), (2.5), and (2.11)-(2.14)

$$|R^A_\omega|\leqq\tilde{R}^A_{\omega(A)} \tag{2.16}$$

in which case the inequality (2.9) is valid at least for all nonsingular block 3-cyclic irreducible H-matrices for which

$$\rho(\tilde{R}^A_{\omega(A)}) < \frac{2r(1-r)}{(1+r)^2} = p(r), \tag{2.17}$$

where, to remind ourselves, $r = \rho(|J^A|)$.

We next turn to the problem of determining the set of all values of r in $(0, 1)$ for which inequality (2.17) is valid. First since A was assumed to be irreducible, obviously

$$|J^A| = \begin{bmatrix} 0 & 0 & C_{13} \\ C_{21} & 0 & 0 \\ 0 & C_{32} & 0 \end{bmatrix} \tag{2.18}$$

is irreducible and so $\tilde{R}^A_{\omega(A)}$ is also irreducible. Moreover, according to a characterization for irreducibility of a nonnegative matrix in canonical form due to Frobenius, 1912 (see Varga [14, Thm. 2.6]), the irreducibility of $|J^A|$ implies that $C = C_{32}C_{21}C_{13}$ is primitive and that

$$\rho(C) = \rho^3(|J^A|) \in \sigma(C). \tag{2.19}$$

It follows that $\tilde{R}^A_{\omega(A)}$ too is primitive since it is an irreducible matrix whose cycle lengths have a greatest common divisor equal to 1.

Continuing, set

$$p_A := \rho(\tilde{R}^A_{\omega(A)}) \tag{2.20}$$

to be the Perron root of $\tilde{R}^A_{\omega(A)}$ and let $u > 0$ be a corresponding Perron vector in which case

$$\tilde{R}^A_{\omega(A)} u = p_A u. \tag{2.21}$$

Now partition the vector u into $u = (x^T, y^T, z^T)^T$ in conformity with the size of the blocks of $\tilde{R}^A_{\omega(A)}$ given in (2.12). Then from (2.12) and (2.21) there results a sequence of substitutions as follows:

$$p_A x = \omega(A)\delta(A)C_{13}z$$

which implies that

$$\begin{aligned} &x = (1/p_A)\omega(A)\delta(A)C_{13}z; \\ &p_A y = \omega(A)\delta(A)C_{21}x + \omega^2(A)\mu^2(A)C_{21}C_{13}z \end{aligned} \tag{2.22}$$

which implies by (2.32) that

$$y = [(1/p_A^2)\omega^2(A)\delta^2(A)C_{21}C_{13} + (1/p_A)\omega^2(A)\mu^2(A)C_{21}C_{13}]z; \tag{2.23}$$

and finally

$$p_A z = \omega^2(A)\delta(A)C_{32}C_{21}x + \omega(A)\delta(A)C_{32}y + \omega^3(A)\mu^2(A)Cz$$

which, by (2.22) and (2.23), reduces to

$$p_A z = \left[\frac{\omega^3(A)\delta^2(A)}{p_A} + \frac{\omega^3(A)\delta^3(A)}{p_A^2} + \frac{\omega^3(A)\delta(A)\mu^2(A)}{p_A} + \omega^3(A)\mu^2(A)\right]Cz. \tag{2.24}$$

Since C is primitive and $z > 0$, (2.19) and (2.24) have the implication that

$$\rho^3(|J^A|) = \frac{p_A^3}{\omega^3(A)\mu^2(A)p_A^2 + [\omega^3(A)\delta^2(A) + \omega^3(A)\delta(A)\mu^2(A)]p_A + \omega^3(A)\delta^3(A)}. \tag{2.25}$$

In passing we mention here that the polynomial relation between p_A and $\rho(|J^A|)$ given by (2.25) is quite distinct, due to different sign cancellation, from the polynomial relation between λ and u given in the paper by Varga, Niethammer and Cai [15, eq. (2.26)].

Let us revert now to the notation $r=\rho(|J^A|)$ which we already introduced following (2.8). If we substitute $\omega(A)=2/(1+r)$ in (2.14) and (2.15) and then substitute the expressions in r thus obtained in (2.25) one deduces, after some simplication, that for $p_A=\rho(\tilde{R}^A_{\omega(A)})$ the following relations hold:

$$p_A^3+B(r)p_A^2+D(r)p_A+E(r)=0, \tag{2.26}$$

where

$$B(r)=-32\frac{r^5}{(1+r)^5}, \tag{2.27}$$

$$D(r)=-32\frac{r^5(1-r)^2}{(1+r)^7}-64\frac{r^6(1-r)}{(1+r)^7}, \tag{2.28}$$

and

$$E(r)=-64\frac{r^6(1-r)^3}{(1+r)^9}. \tag{2.29}$$

Consider the parametric cubic equation

$$\tau^3+B(r)\tau^2+D(r)\tau+E(r)=0,\ r\in(0,1). \tag{2.30}$$

By (2.26) for each H-matrix A with $r=\rho(|J^A|)$, the Perron root p_A of $\tilde{R}^A_{\omega(A)}$ is a positive real root of (2.30) whose size, in view of the discussion leading to (2.17), we wish to compare with the quantity $p(r)=2r(1-r)/(1+r)^2$. Let us first determine the possible values $r\in(0,1)$ for which $p(r)=p_A$. For this purpose let us substitute the expression for $p(r)$ in (2.30). One obtains after some simplifications, the following *quartic* equation in r:

$$f(r):=17r^4+18r^3-2r-1=(r+1)(17r^3+r^2-r-1)=0. \tag{2.31}$$

It is simple to ascertain using the discriminant test for the roots of a cubic (see [13, pp. 103-104]) that $17r^3+r^2-r-1$ has one positive real root $r_0\in(0,1)$ and two complex conjugate roots. It follows that the quartic $f(r)$ given in (2.31) has precisely one root r_0 in the interval $(0,1)$ and hence r_0 is the only value of r in $(0,1)$ for which $p(r)=p_A$. Computations show that

$$r_0\simeq .4181928. \tag{2.32}$$

In order to determine more precisely the relationship between p_A and $p(r)$ at least in an interval containing $(0,r_0)$ we next prove the following auxiliary lemma

LEMMA 2.1. (i) *Of the three branches of solutions to* (2.30) *only one of them* x_r *is an increasing function at least in the interval* $(0,5/7)$.

(ii) *The function* $p(r)=2r(1-r)/(1+r)^2$ *is positive and concave in* $(0,1)$ *and it takes its maximum value in the interval at* $r=1/3$.

Proof. The proof of part (ii) is essentially a freshman's calculus exercise and so only part (i) of the Lemma will be proved here.

First using expressions (2.27)-(2.29) for $B(r)$, $D(r)$ and $E(r)$ one can verify the existence of values of r in $(0,1)$, e.g., $r=0.25$ for which the discriminant of (2.30) is positive and hence for the values of r in $(0,1)$ the Perron root of $\tilde{R}^A_{\omega(A)}$, which in this lemma we denote by x_r, is the only branch of solution to (2.30) which is positive throughout the interval (0, 1). To complete the proof of the part (i), we shall therefore show that

$$\frac{dx_r}{dr} \geqq 0 \quad \forall r \in (0, 5/7). \tag{2.33}$$

Implicit differentiation of

$$x_r^3 + B(r)x_r^3 + D(r)x_r + E(r) = 0 \tag{2.34}$$

with respect to r yields that

$$\frac{dx_r}{dr} = -\frac{B'(r)x_r^2 + D'(r)x_r + E'(r)}{3x_r^2 + 2B(r)x_r + D(r)}. \tag{2.35}$$

Hence, on multiplying both sides of (2.35) by $(1/x_r) > 0$ one obtains that

$$\frac{1}{x_r}\frac{dx_r}{dr} = -\frac{B'(r)x_r^2 + D'(r)x_r + E'(r)}{3x_r^3 + 2B(r)x_r^2 + D(r)x_r}. \tag{2.36}$$

Consider the denominator on the right-hand side of (2.36):

$$\begin{aligned} &3x_r^3 + 2B(r)x_r^2 + D(r)x_r \\ &\quad = 3[x_r^3 + B(r)x_r^2 + D(r)x_r + E(r)] - B(r)x_r^2 - 2D(r)x_r - 3E(r) \\ &\quad = -B(r)x_r^2 - 2D(r)x_r - 3E(r) > 0 \quad \forall r \in (0, 1). \end{aligned} \tag{2.37}$$

We mentioned that the last equality in (2.37) follows by (2.27)-(2.29), (2.34), and because $x_r > 0$. To prove (2.33), it therefore remains to show that at least for $r \in (0, 5/7)$ the numerator of (2.36) is negative. Indeed,

$$B'(r) = -160\frac{r^4}{(1+r)^6} < 0 \quad \forall r \in (0, 1),$$

$$D'(r) = -\frac{32r^4(5 - 2r - 7r^2)}{(1+r)^8} < 0 \quad \forall r \in (0, 5/7),$$

and

$$E'(r) = -\frac{192r^5(2 - 8r + 10r^2 - r^3)}{(1+r)^{10}} < 0 \quad \forall r \in (0, 1).$$

This completes the proof.

The situation is quite clear now:

$$p_A = \rho(\tilde{R}^A_{\omega(A)}) < p(r) = \frac{2r(1-r)}{(1+r)^2} \quad \text{whenever } \rho(|J^A|) \in (0, r_0). \tag{2.38}$$

This is born by the graph in Fig. 1. We note that from Fig. 1 it appears that $\rho(\tilde{R}^A_{\omega(A)})$ is an increasing function of $r = \rho(|J^A|)$ throughout the interval (0, 1) a fact which we have not been able to prove here, but which we did not require either. We further mention that our computations show that the discriminant of the cubic (2.30) is positive for r in the range $0 < r < .6$.

Lemma 2.1 and (2.38) together with the analysis presented in (2.1)-(2.17) yield the following result:

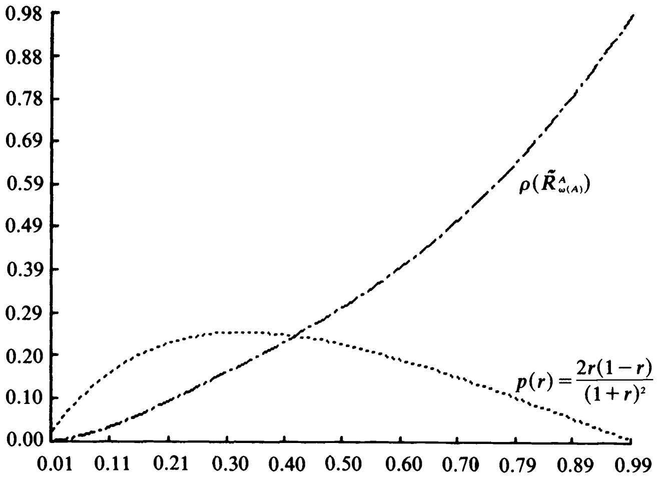


FIG. 1

THEOREM 2.2. *Let A be an $n\times n$ complex nonsingular irreducible* 3*-cyclic H-matrix. If $\rho(|J^A|)\in(0, r_0)$, where r_0 is the unique root of the cubic* (1.12) *in the interval* (0, 1), *then*

$$\rho(S^A_{\omega(A)})\leqq\rho(\tilde{S}^A_{\omega(A)})<\omega(A)-1\leqq\rho(L^A_{\omega(A)}). \tag{2.39}$$

As indicated previously, r_0 is roughly given (2.32). It is evident now that because of the strict inequality in (2.4) a neighborhood of dominant convergence of the SSOR method over the SOR method, in the sense defined in § 1, exists for each H-matrix A with $\rho(|J^A|)\in(0, r_0)$. This proves that the statement of Corollary 1.2 is valid.

Figure 2, where

$$\tilde{\tilde{S}}^A_{\omega(A)}=(1-\omega(A))^2I+\tilde{R}^A_{\omega(A)},$$

illustrates that the inequalities in (2.39) between $\rho(S^A_{\omega(A)})$ and $\rho(L^A_{\omega(A)})$ remain valid considerably beyond r_0, but it appears that different tools need to be developed to examine the relationships between $\rho(S^A_{\omega(A)})$ and $\rho(L^A_{\omega(A)})$ outside the interval $(0, r_0)$.

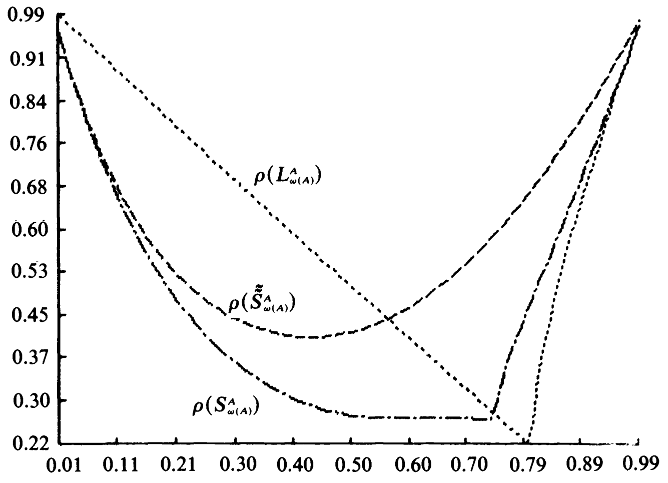


FIG. 2

Finally, in view of Observation 1.1 and Corollary 1.2 we raise here the following question:

Suppose A is an $n \times n$ nonsingular M-matrix such that $\rho(J^A) \in (0, r_0)$. Does the inequality

$$\rho(S_\omega^A) \geqq \rho(L_\omega^A)$$

hold for all $\omega \in (0, \omega(A))$?

Acknowledgments. The author would like to thank Drs. Nieves McNulty and David Richman for very helpful discussions. The author is most indebted to Professor Apostolos Hadjidimos for detecting an error in one of the constructions in the original manuscript. Finally, he wishes to thank Professor R. J. Plemmons and R. S. Varga for additional useful comments.

REFERENCES

[1] G. ALEFELD AND R. S. VARGA, *zur Konvergenz des symmetrischen Relaxionsverfahrens*, Numer. Math., 25 (1976), pp. 291-295.

[2] O. AXELSSON, *A survey of preconditioned iterative methods for linear systems of algebraic equations*, BIT, 25 (1985), pp. 166-187.

[3] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[4] C. CSORDAS AND R. S. VARGA, *Comparisons of regular splittings of matrices*, Numer. Math., 44 (1984), pp. 23-35.

[5] W. KAHAN, *Gauss-Seidel methods for solving large systems of linear equations*, Ph.D. thesis, Univ. of Toronto, Toronto, Ontario, Canada, 1958.

[6] U. KULISCH, *Uber regular zerlegungen von Matrizen und einige Andwendungen*, Numer. Math., 11 (1968), pp. 444-449.

[7] V. A. MILLER AND M. NEUMANN, *A note on comparison theorems for nonnegative matrices*, Numer. Math., 47 (1985), pp. 427-434.

[8] A. NEUMAIER AND R. S. VARGA, *Exact convergence and divergence domains for the symmetric successive overrelaxation iterative method applied to H-matrices*, Linear Algebra Appl., 58 (1984), pp. 261-272.

[9] M. NEUMANN, *The Kahan* SOR *convergence bound for nonsingular and irreducible M-matrices*, Linear Algebra Appl., 39 (1981), pp. 205-222.

[10] ———, *On bounds for the convergence of the* SSOR *method for H-matrices*, Linear and Multilinear Algebra, 15 (1984), pp. 13-21.

[11] ———, *On bounds for the convergence of the* SSOR *Method for H-matrices*, preprint, 1981.

[12] M. NEUMANN AND R. S. VARGA, *On the sharpness of some upper bounds for the spectral radii of* S.O.R. *iteration matrices*, Numer. Math., 35 (1980), pp. 69-79.

[13] S. M. SELBY, *Standard Mathematical Tables*, 18th ed., The Chemical Rubber Co., Cleveland, 1970.

[14] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[15] R. S. VARGA, W. NIETHAMMER AND D. Y. CAI, *p-cyclic matrices and the symmetric successive overrelaxation method*, Linear Algebra Appl., 58 (1984), pp. 425-439.

[16] Z. WOŹNICKI, *Two-sweep iterative method for solving large linear systems and their application to the numerical solution of multi-group multi-dimensional neutron diffusion equations*, Ph.D. dissertation, Report No. 1447/CTFRIBET/PM/A, Inst. Badan Jadrowych, Warzwa, 1973.

[17] D. M. YOUNG, *Iterative Solutions of Large Linear Systems*, Academic Press, New York, 1971.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 4, October 1986

006

FREDMAN-KOMLÓS BOUNDS AND INFORMATION THEORY*

JÁNOS KÖRNER†

Abstract. Fredman and Komlós have applied an interesting information-theoretic lemma to two problems in combinatorics. They have derived good lower bounds on the minimum size of a family of partitions of an n-element set into at most b classes such that all the subsets (respectively, pairs of subsets) of a certain kind are "separated" by at least one partition in the family.

Our aim is to show that the Fredman-Komlós lemma is a special case of a simple inequality between entropies of graphs. The general inequality enables us to handle more problems on separating partition systems. Part of the problems relate to hashing.

Key words. perfect hashing, graph covering, graph entropy

AMS(MOS) subject classifications. Primary 68E10; secondary 05C70, 94A15

1. Introduction. In a recent paper Fredman and Komlós [4] have applied an interesting information-theoretic technique to two problems in combinatorics. We shall extend and simplify their method in order to be able to treat a large class of problems on separating partition systems in a perspicuous manner.

The key notion in our approach is a functional, introduced for arbitrary graphs by Kőrner [5]. We shall point out later on that this notion is implicitly contained in the basic lemma of Fredman and Komlós in the case of a graph consisting of a complete subgraph plus an isolated point.

In § 2 we shall recall the definition and some basic properties of graph entropy. Furthermore, we shall prove our basic inequality. To illustrate the advantage of our generalization, we shall rederive the Fredman-Komlós result on perfect hash functions [4] in a short and elegant manner in § 3. In § 4 we shall treat a related problem, nearly perfect hash functions, that seems to be untractable by the original Fredman-Komlós technique. In the final § 5 of this paper we shall discuss the tightness of our bounds and explain the relation of our approach to that in [4]. Also, we shall point out that graph entropy is a better estimate for the Fredman-Komlós "content" of a graph than their Theorem 1.

No information-theoretic prerequisites are needed. However, for more details on our terminology and the basic notions of information theory we refer the reader to the books by Csiszár and Körner [2] or McEliece [7].

All logarithms are to the base 2. For the reader's convenience we recall that the entropy $H(P)$ of the probability distribution P on the finite set $\mathbf{X}$ is given by

$$H(P) = -\sum_{x\in\mathbf{X}} P(x)\log P(x).$$

The entropy $H(P)$ of a random variable (RV) X with values in $\mathbf{X}$ and distribution $P\{X=x\}=P(x)$ is defined as $H(X)=H(P)$. Given the RV's X and Y with respective finite ranges $\mathbf{X}$ and $\mathbf{Y}$, the mutual information $I(X\wedge Y)$ between X and Y is

$$I(X\wedge Y)=H(X)+H(Y)-H(X,Y)$$

where $H(X, Y)$, the joint entropy of (X, Y) is simply the entropy of the RV (X, Y) that takes its values in $\mathbf{X}\times\mathbf{Y}$.

* Received by the editors February 6, 1985, and in revised form January 13, 1986. Part of this research was done while the author was visiting Trieste University, Italy.

† Mathematical Institute of the Hungarian Academy of Sciences, 1053 Budapest, Hungary.

The term *graph* means a simple graph, i.e., one with undirected edges, without multiple edges and loops. An *independent set* of the graph G is a subset of the vertex set in which no pairs of vertices are connected by an edge of G. The maximum cardinality of such a set, denoted by $\alpha(G)$, is called the *independence number* of G. A *coloring* of G is a partition of the vertex set into independent sets called color classes. The *chromatic number* $\gamma(G)$ of the graph G is the minimum number of classes in any coloring of G. We shall denote by $\mathbf{V}(G)$ the set of vertices of G, while $\mathbf{E}(G)$ will refer to the set of its edges. Note that the elements of $\mathbf{E}(G)$ are unordered pairs of elements of $\mathbf{V}(G)$. The term *subgraph* will always refer to induced subgraphs. The inclusion $F \subset G$ means that F is an induced subgraph of G. For the unexplained terminology concerning graphs, cf. Berge [1].

2. Graph entropy. Let us consider a graph G with vertex set $\mathbf{X}$ and an arbitrary probability distribution P on the set $\mathbf{X}$. We call the couple (G, P) a *probabilistic graph.* Graph entropy is a nonnegative valued functional on probabilistic graphs, introduced in [5]. Reversing the role of the definition and characterization in that paper, we shall define graph entropy as follows:

DEFINITION 1. Let $\mathcal{A}$ be the family of all the maximal independent sets of G. A distribution Q will be said to belong to the family $\mathcal{P}(G)$ of distributions on $\mathbf{X}\times\mathcal{A}$ if it is concentrated on those elements $(x, y)\in\mathbf{X}\times\mathcal{A}$ for which the vertex x is contained in the independent set y. Set

$$H(G, P)\triangleq\min\{I(X\wedge Y);\ P_X=P,\ P_{XY}\in\mathcal{P}(G)\}.$$

$H(G, P)$ is the *graph entropy* of (G, P).

The key element in our machinery is a simple inequality for graph entropies.

DEFINITION 2. Let the graphs F and G have the same set of vertices. The *union graph* $F\cup G$ is defined by the relations

$$\mathbf{V}(F\cup G)=\mathbf{V}(F)=\mathbf{V}(G),\qquad \mathbf{E}(F\cup G)=\mathbf{E}(F)\cup\mathbf{E}(G).$$

We have our key

LEMMA 1. *For every probability distribution P on* $\mathbf{V}(F)=\mathbf{V}(G)$,

$$H(F\cup G, P)\leqq H(F, P)+H(G, P).$$

Proof. We define the triple of RV's X, Y, Z as follows. Let X have distribution P, let P_{XY} attain the minimum in the definition of $H(F, P)$, and, likewise, let P_{XZ} attain $H(G, P)$. Further, let Y and Z be conditionally independent given X. Clearly, the maximal independent sets of $F\cup G$ are exactly the pairwise intersections of the maximal independent sets of F and G. Hence the pair of RV's X, Y can be regarded to take its values in the maximal independent sets of $F\cup G$. Obviously, the joint distribution of X and YZ is contained in $\mathcal{P}(F\cup G)$. Hence

$$H(F\cup G,\ P)\leqq I(X\wedge YZ). \tag{1}$$

Now, by the well-known nonnegativity of mutual information, (cf. [2, Lemma 1.3.2]),

$$I(X\wedge YZ)\leqq I(X\wedge YZ)+I(Y\wedge Z). \tag{2}$$

But, by our definition of mutual information, the right-hand side of this inequality equals

$$H(X)+H(Y)+H(Z)-H(XYZ),$$

an expression symmetric in X, Y and Z. Hence (2) can be rewritten as

$$I(X\wedge YZ)\leqq I(X\wedge Y)+I(XY\wedge Z)=I(X\wedge Y)+I(X\wedge Z),$$

where the last equality follows from the fact that Y and Z are conditionally independent given X. Comparing this with (1) and recalling that $I(X \wedge Y)$ and $I(X \wedge Z)$ achieve $H(F, P)$ respectively, $H(G, P)$, the statement of the lemma follows. □

We shall return to the problem of equality later on. By an iterative application of the lemma we get

COROLLARY 1. *For every probability distribution P on $\mathbf{V}(G_1) = \mathbf{V}(G_2) = \cdots = \mathbf{V}(G_k)$*

$$H\left(\bigcup_{i=1}^{k} G_i, P\right) \leqq \sum_{i=1}^{k} H(G_i, P).$$

We add a simple observation. Let $\mathbf{A}$ be an independent set of the graph G. Define the graph $G(\mathbf{A})$ as the result of the operation of replacing $\mathbf{A}$ by a single vertex adjacent to precisely those vertices of G which were adjacent to at least one element of $\mathbf{A}$. Let the distribution $P^{\mathbf{A}}$ be obtained from P by assigning to the single vertex that replaces $\mathbf{A}$ the total probability $P(\mathbf{A})$. It easily follows from the definition of graph entropy that

LEMMA 2.

$$H(G, P) \leqq H(G(\mathbf{A}), P^{\mathbf{A}}).$$

It is clear that

$$H(G, P) \leqq \log \gamma(G).$$

Further,

LEMMA K (Körner [5]). *Let G be a graph all the connected components G_i of which are complete subgraphs. Further, set*

$$P_i(x) = P(x)[P(G_i)]^{-1}, \qquad x \in G_i.$$

Then

$$H(G, P) = \sum_i P(G_i) H(P_i).$$

In § 4 we shall prove a generalization of this lemma.

The above is all we need to rederive the Fredman-Komlós bound in a few lines, as it will be done in the next section. However, it would be inappropriate not to dwell somewhat more on this notion so as to provide some intuitive picture. I will try to explain, using my original definition, that graph entropy is a probabilistically refined concept of the chromatic number of a graph.

Given the graph G with vertex set $\mathbf{X}$, let us define the nth power of G^n of G as follows. The vertex set of G^n is $\mathbf{X}^n$. Two vertices, $\mathbf{x}$ and $\mathbf{x}'$ of G^n are adjacent (connected by an edge) in G^n if for at least one of their n coordinates the corresponding vertices of the two sequences are connected by an edge in G. Intuitively, we might think of G as the graph of distinguishability between the elements of $\mathbf{X}$. Then, the two endpoints of an edge are always considered to be distinguishable, while the remaining pairs of vertices might be confused. The graph G^n is the natural extension of this relation to sequences of length n of elements of $\mathbf{X}$.

Consider an arbitrary probability distribution P on the set $\mathbf{X}$. We denote by P^n the product extension of P to $\mathbf{X}^n$, i.e., for $\mathbf{x} = x_1 x_2 \cdots x_n$ we write

$$P^n(\mathbf{x}) \triangleq \prod_{i=1}^{n} P(x_i).$$

We define the ε-chromatic number $\gamma_\varepsilon(G, P)$ to be the minimum of the chromatic numbers of the "large" subgraphs of G;

$$\gamma_\varepsilon(G, P) \triangleq \min \{\gamma(F);\ P(F) \geqq 1 - \varepsilon,\ F \subset G\}.$$

In this definition, $P(F)$ is the total probability of the vertices of F. It is easy to see that for every fixed ε, the quantity $\gamma_\varepsilon(G^n, P^n)$ is exponential in n. We recall

THEOREM K (Körner [5]). *For every* $\varepsilon \in (0, 1)$

$$\lim_{n\to\infty} \frac{1}{n} \log \gamma_\varepsilon(G^n, P^n) = H(G, P).$$

Remark. As we have stressed before, this characterization theorem is not used in our proof of the Fredman-Komlós result. Originally, the author has obtained it in an esoteric information theoretic context. In our original paper introducing graph entropy, it was defined as the left-hand limit above. Theorem K served to prove that this limit exists, is independent of $\varepsilon \in (0, 1)$ and can be described by the formula we are using here as the definition. We have reversed the function of the two mainly to point out the brevity and the elegance of the proof of Theorem FK.

3. Perfect hash functions. Consider the n-element set $\mathbf{X}$. We shall say that a function $f:\mathbf{X}\to\mathbf{B}$ *separates* the set $\mathbf{A}\subset\mathbf{X}$ if f takes a different value on every element of $\mathbf{A}$. Let f_π, $\pi\in\Pi$ be a family of mappings of the set $\mathbf{X}$ into a set $\mathbf{B}$. Set $b = |\mathbf{B}|$. The family $\{f_\pi\}_{\pi\in\Pi}$ is said to be a (b, k)*-family of perfect hash functions* for $\mathbf{X}$ if every k-element subset $\mathbf{A}\subset\mathbf{X}$ is separated by at least one function f_π, $\pi\in\Pi$. Let us denote by $Y(b, k, n)$ the minimum size of any (b, k)-family of perfect hash functions for $\mathbf{X}$. We are interested in the asymptotics of $Y(b, k, n)$ for every fixed b and k.

We shall define the functions f_π through partitions of $\mathbf{X}$, assigning a different element of $\mathbf{B}$ to the different classes of a partition in an arbitrary manner. Selecting each partition equiprobably, independently from one another among all possible equipartitions of $\mathbf{X}$ into b classes, one easily sees, cf. [4], that

$$Y(b, k, n) \lesssim \frac{b^k}{b^{\underline{k}}} \cdot k \cdot \log n \tag{3}$$

as n goes to infinity, while b and k remain fixed. Here and in the sequel,

$$b^{\underline{k}} \triangleq \prod_{i=0}^{k-1} (b-i).$$

Next we prove that

THEOREM FK (Fredman-Komlós [4]).

$$Y(b, k, n) \gtrsim \frac{b^{k-1}}{b^{\underline{k-1}}} \frac{\log n}{\log (b-k+2)}, \tag{4}$$

asymptotically in n, for every fixed b and k.

In order to rederive this result, let us first observe the obvious

LEMMA FK. *The number of k-element sets in* $\mathbf{X}$ *separated by any given* $f:\mathbf{X}\to\mathbf{B}$ *is upper bounded asymptotically by*

$$\frac{b^{\underline{k}}}{b^k} \cdot \binom{n}{k}.$$

Proof of Theorem FK. Let us define a graph G the vertices of which are pairs $(x, \mathbf{C})$ where $x\in\mathbf{X}$ and $\mathbf{C}$ is a $(k-2)$-element subset of $\mathbf{X}$ that does not contain x. Let $(x, \mathbf{C})$ and $(x', \mathbf{C}')$ be adjacent in G if and only if $x\neq x'$, $\mathbf{C}=\mathbf{C}'$. Clearly, for every $\mathbf{C}\subset\mathbf{X}$ with $|\mathbf{C}| = k-2$, G has a connected component which is a complete subgraph on $n-k+2$ elements. Hence, letting P be the equidistribution on the vertex set of G and noticing that these subgraphs exhaust the vertex set of G, we have, by Lemma K,

$$H(G, P) = \log (n-k+2). \tag{5}$$

Now, let $f_\pi : \mathbf{X} \to \mathbf{B}$, $\pi \in \Pi$ be an arbitrary (b, k)-family of perfect hash functions for $\mathbf{X}$. With every function f_π we shall associate a graph G_π having the same vertex set as G. Let $(x, \mathbf{C})$ and $(x', \mathbf{C}')$ be adjacent in G_π if and only if

(i) they are adjacent in G,

(ii) the k-element set obtained from $\mathbf{C} = \mathbf{C}'$ by adding the elements x and x' is separated by f_π.

Then the fact that f_π, $\pi \in \Pi$ is a (b, k)-family implies that G is the union of the graphs G_π, $\pi \in \Pi$ in the sense of Definition 2. Hence, by Corollary 1, we get

$$H(G, P) \leqq \sum_{\pi \in \Pi} H(G_\pi, P).$$

If, furthermore, $\{f_\pi\}_{\pi \in \Pi}$ is an optimal (b, k)-family, the last inequality yields

$$H(G, P) \leqq Y(b, k, n) \max_{\pi \in \Pi} H(G_\pi, P). \tag{6}$$

It remains to upper bound $H(G_\pi, P)$. It is clear from the definition of G_π that the points $(x, \mathbf{C})$ for which the set $\mathbf{C} \cup \{x\}$ is not separated by f_π are all isolated points in G_π. Further, if $(x, \mathbf{C})$ and $(x', \mathbf{C}')$ are adjacent in G_π, then $\mathbf{C} = \mathbf{C}'$ and $f_\pi(x) \neq f_\pi(x')$. Moreover, all the pairs $(x, \mathbf{C})$ for which $(f_\pi(x), f_\pi(\mathbf{C}))$ is the same pair of sets are in a corresponding single independent set of G_π. Iteratively applying the contraction operation of Lemma 2 to these independent sets, we obtain a graph that contains a set of nonisolated points having total probability at most

$$\frac{b^{\underline{k-1}}}{b^{k+1}},$$

(cf. Lemma FK), the subgraph of which has the property that its connected components are complete subgraphs of $b - k + 2$ vertices each. Let us denote this new graph by (G'_π, P'). By Lemma 2,

$$H(G_\pi, P) \leqq H(G'_\pi, P'). \tag{7}$$

On the other hand, we have seen that G'_π satisfies the conditions of Lemma K, and hence

$$H(G'_\pi, P') \leqq \frac{b^{\underline{k-1}}}{b^{k-1}} \log (b - k + 2).$$

Comparing this inequality with (5)-(7), we obtain the statement of the theorem. □

In the last section of this paper we shall return to the question of the tightness of this bound. Further, we shall explain in what sense this is a streamlined version of the Fredman-Komlós proof. For an explanation of the hitherto mysterious title of this section we refer the reader to [4].

4. Nearly-perfect hash functions. Lemma 1 applies to a wealth of problems in which a family of partitions must "separate" subsets or t-tuples of subsets of a given ground set. Not all of these problems are however tractable by the original Fredman-Komlós technique. Crucial to their approach is namely the reliance on graphs all the connected components of which are complete. We shall see in what follows that not all the problems of the above type can be described in terms of such graphs. To illustrate this phenomenon, we introduce a slight variation into the previous problem.

Let us have $|\mathbf{X}| = n$, $|\mathbf{B}| = b$. We shall say that a function $f : \mathbf{X} \to \mathbf{B}$ *nearly separates* the set $\mathbf{A} \subset \mathbf{X}$ if f takes $|\mathbf{A}| - 1$ different values on $\mathbf{A}$. The family $\{f_\pi\}_{\pi \in \Pi}$ of mappings $f : \mathbf{X} \to \mathbf{B}$ is said to be a *(b, k)-family of nearly-perfect hash* functions for $\mathbf{X}$ if every k-element subset of $\mathbf{X}$ is nearly separated by at least one function f_π, $\pi \in \Pi$. Let us denote by $Z(b, k, n)$ the minimum size of any such family for $\mathbf{X}$. We are interested in the asymptotics of this quantity for any fixed b and k.

Let $F(b, k, l, n)$ denote the maximum fraction of k-element subsets of $\mathbf{X}$ on which a function $f:\mathbf{X}\to\mathbf{B}$ takes precisely l different values. Since it is easily seen that $F(b, k, l, n)$ is achieved by equipartitions of $\mathbf{X}$, we have the asymptotic inequality

$$(8)\qquad F(b,k,l,n)\lesssim\binom{b}{l}\binom{n}{b}^{l}\binom{\frac{n}{b}l}{k-l}\cdot\left[\binom{k}{l}\binom{n}{k}\right]^{-1}\sim\frac{b^{\underline{l}}}{b^k}l^{k-l},$$

for every fixed b, k and l. In particular,

$$F(b,k,n)\triangleq F(b,k,k-1,n)\lesssim\frac{b^{\underline{k-1}}}{b^k}(k-1).$$

A standard random selection argument yields

LEMMA 3.

$$Z(b,k,n)\lesssim\frac{b^k}{b^{\underline{k-1}}}\cdot\frac{k}{k-1}\cdot\log n,$$

asymptotically, for fixed b and k.

Proof. As in the previous problem, we shall select partitions which will be used to define the functions in an obvious way. Let us choose our partitions with equal probabilities and independently from one another among all the possible equipartitions of $\mathbf{X}$ into b classes. After m partitions have been selected, the probability that there is at least one k-element subset of $\mathbf{X}$ which is not nearly separated by at least one of our partitions has the obvious upper bound (cf. (8))

$$\left[1-(k-1)\frac{b^{\underline{k-1}}}{b^k}\right]^m\binom{n}{k}.$$

If this expression is strictly less than 1, then there is at least one family consisting of m partitions that has the desired properties. Hence

$$m\sim k\cdot\log n\left\{-\log\left[1-(k-1)\frac{b^{\underline{k-1}}}{b^k}\right]\right\}^{-1}\lesssim\frac{b^k}{b^{\underline{k-1}}}\cdot\frac{k}{k-1}\cdot\log n.\qquad\square$$

The counterpart of the last lemma is a generalization of the Fredman-Komlós type lower bounds. In order to prove it, we need a generalization of Lemma K.

LEMMA K*. *Let the connected components of the graph G be the subgraphs G_i. Further, set*

$$P_i(x)=P(x)[P(G_i)]^{-1},\qquad x\in G_i.$$

Then

$$H(G,P)=\sum_i P(G_i)H(G_i,P_i).$$

Proof. The inequality

$$H(G,P)\geqq\sum_i P(G_i)H(G_i,P_i)$$

follows from Theorem K. In fact, let $G(n)$ be a graph with vertex set $[\mathbf{V}(G)]^n$. Let the vertices x and x' of $G(n)$ be connected by an edge in $G(n)$ if they are adjacent in G^n

and all their coordinates belong to the same subgraphs G_i. Let us fix some $\varepsilon \in (0, 1)$. It is easily seen that

$$\liminf_{n\to\infty} \frac{1}{n} \log \gamma_\varepsilon(G(n), P^n) \geqq \sum_i P(G_i) H(G_i, P_i).$$

As $\mathbf{E}(G(n)) \subset \mathbf{E}(G^n)$, it is also clear that

$$\gamma_\varepsilon(G^n, P^n) \geqq \gamma_\varepsilon(G(n), P^n).$$

Comparing this with the previous inequality, our first assertion follows by the definition of $H(G, P)$ and Theorem K. To prove inequality in the other direction, let G_i^* be the graph with vertex set $\mathbf{V}(G)$ obtained from G_i by adding the remaining points of $\mathbf{V}(G)$ as isolated points. By Lemma 1,

$$H(G, P) \leqq \sum_i H(G_i^*, P).$$

It remains to see that

$$H(G_i^*, P) \leqq P(G_i) H(G_i, P_i).$$

To this end, let P_{XY} achieve $H(G_i, P_i)$ for some fixed i. To any maximal independent set $\mathbf{A}$ in G_i there corresponds a maximal independent set $\mathbf{A}^*$ in G_i^* obtained by adding to $\mathbf{A}$ all the isolated points in G_i^*. Let us define the joint distribution of the RV's X^*, Y^* as follows. Set $P_{X^*} \triangleq P$. Let the conditional distributions of Y^* given the various values of X^* be

$$P_{Y^*|X^*}(\mathbf{A}^*|x) \triangleq \begin{cases} P_{Y|X}(\mathbf{A}|x) & \text{if } x \in \mathbf{V}(G_i), \\ P_Y(\mathbf{A}) & \text{else.} \end{cases}$$

An easy computation shows that

$$I(X^* \wedge Y^*) = H(Y^*) - \sum_{x \in \mathbf{V}(G)} P(x) H(Y^*|X^* = x)$$

where $H(Y^*|X^* = x)$ is the entropy of the probability distribution $P_{Y^*|X^*}(\cdot|x)$. Hence, by definition,

$$\begin{aligned} I(X^* \wedge Y^*) &= \sum_{x \in \mathbf{V}(G_i)} P(x)[H(Y^*) - H(Y^*|X^* = x)] \\ &= \sum_{x \in \mathbf{V}(G_i)} P(x)[H(Y) - H(Y|X = x)] = P(G_i) I(X \wedge Y). \end{aligned}$$

Noting that $H(G_i^*) \leqq I(X^* \wedge Y^*)$ and $I(X \wedge Y) = H(G_i, P_i)$, we can complete the proof. □

Remark. It is possible to prove this lemma directly, without using Theorem K.

Now we are ready to establish

THEOREM 1. *For every fixed b and k, asymptotically in n,*

$$Z(b, k, n) \gtrsim \frac{b^{k-3}}{b^{\underline{k-3}}} \frac{\log n}{\log(b - k + 5)}. \tag{9}$$

Proof. Let G be a graph the vertices of which are pairs of subsets $(\mathbf{A}, \mathbf{C})$ of $\mathbf{X}$ such that $|\mathbf{A}| = 2$, $|\mathbf{C}| = k - 4$. The vertices $(\mathbf{A}, \mathbf{C})$ and $(\mathbf{A}', \mathbf{C}')$ are adjacent in G if and only if $\mathbf{C} = \mathbf{C}'$ and the sets $\mathbf{A}, \mathbf{A}', \mathbf{C}$ are pairwise disjoint. Notice that the last condition implies $|\mathbf{A} \cup \mathbf{A}' \cup \mathbf{C}| = k$. Let P be the equidistribution over the vertices of G.

Then, clearly, the graph G has $\binom{n}{k-4}$ connected components all of which are isomorphic. Let G' be such a connected component and let P' be the equidistribution over its set of vertices. By Lemma K*

$$H(G, P) = H(G', P').$$

Further, by the Erdös-Ko-Rado theorem [3],

$$\alpha(G') \leqq n-k+3.$$

By the last two relations we see that

$$H(G, P) = H(G', P') \geqq \log \frac{|\mathbf{V}(G')|}{\alpha(G')} \geqq \log \frac{\binom{n-k+4}{2}}{n-k+3} \gtrsim \log n. \tag{10}$$

(The first inequality in (10) is obvious. Nevertheless, its proof will be given in the last section, cf. (14).)

Let f_π, $\pi \in \Pi$ be a family of mappings from $\mathbf{X}$ to $\mathbf{B}$ that achieves the minimum in the definition of $Z(b, k, n)$. For every $\pi \in \Pi$, let us define the graph G_π with vertex set equal to $\mathbf{V}(G)$ as follows. Let $(\mathbf{A}, \mathbf{C})$ and $(\mathbf{A}', \mathbf{C}')$ be connected by an edge in G_π if and only if

(i) they are adjacent in G,

(ii) $\mathbf{A} \cup \mathbf{A}' \cup \mathbf{C}$ is nearly separated by f_π.

Since the family f_π, $\pi \in \Pi$ is a (b, k)-family of nearly-perfect hash functions for $\mathbf{X}$, we have

$$G = \bigcup_{\pi \in \Pi} G_\pi,$$

whence, by Corollary 1, we obtain

$$H(G, P) \leqq \sum_{\pi \in \Pi} H(G_\pi, P) \leqq Z(b, k, n) \cdot \max_{\pi \in \Pi} H(G_\pi, P). \tag{11}$$

As in the previous theorem, it remains to upper bound $H(G_\pi, P)$.

We shall refer to the different values of f_π as "colors." Let $f_\pi(\mathbf{A})$ denote the set of all the different colors f_π takes on the various elements of $\mathbf{A}$. Then, $f_\pi(\mathbf{A}) \subset \mathbf{B}$. Clearly, if $(\mathbf{A}, \mathbf{C})$ is not an isolated point, then either

(a) $|f_\pi(\mathbf{C})| = k-5, \quad |f_\pi(\mathbf{A} \cup \mathbf{C})| = k-3,$

(b) $|f_\pi(\mathbf{C})| = k-4, \quad |f_\pi(\mathbf{A} \cup \mathbf{C})| = k-3,$

(c) $|f_\pi(\mathbf{C})| = k-4, \quad |f_\pi(\mathbf{A} \cup \mathbf{C})| = k-2.$

Let us denote the sets comprising all the vertices of G_π with one of the above properties by $\mathbf{X}_a$, $\mathbf{X}_b$ and $\mathbf{X}_c$, respectively. As in the previous problem, we can see that all the $(\mathbf{A}, \mathbf{C})$ for which $(f_\pi(\mathbf{A}), f_\pi(\mathbf{C}))$ is the same set pair of colors form an independent set of G_π. Iteratively contracting each of these sets into a different single vertex we obtain a graph (G'_π, P'). By Lemma 2,

$$H(G_\pi, P) \leqq H(G'_\pi, P'). \tag{12}$$

Let us denote by $\mathbf{X}'_a$, $\mathbf{X}'_b$ resp. $\mathbf{X}'_c$ the contracted versions of $\mathbf{X}_a$, $\mathbf{X}_b$ and $\mathbf{X}_c$, respectively. Clearly,

$$P'(\mathbf{X}'_a) = P(\mathbf{X}_a), \quad P'(\mathbf{X}'_b) = P(\mathbf{X}_b), \quad P'(\mathbf{X}'_c) = P(\mathbf{X}_c). \tag{13}$$

It is easy to see that the subgraph of G'_π defined by the vertex set $\mathbf{X}'_a$ has chromatic number at most $b-k+5$. Clearly, the vertices of $\mathbf{X}'_b$ form a single independent set in G'_π. Further, the chromatic number of the subgraph on $\mathbf{X}'_c$ equals the minimum number of colors needed to color the edges of the complete graph on $b-k+4$ vertices so that adjacent edges get different colors. As is well known, cf. [1, Thm. 1, p. 249], this number is at most $b-k+4$.

Let $G_\pi(a)$ denote the subgraph of G'_π having vertex set $\mathbf{X}'_a$ and let $G_\pi(b, c)$ denote its subgraph with vertex set $\mathbf{X}'_b \cup \mathbf{X}'_c$. The above analysis shows that

$$\gamma(G_\pi(a)) \leqq b-k+5, \qquad \gamma(G_\pi(b, c)) \leqq \gamma(G_\pi(c))+1 \leqq b-k+5.$$

By Lemma K*, this implies

$$H(G'_\pi, P') \leqq P'(\mathbf{X}'_a \cup \mathbf{X}'_b \cup \mathbf{X}'_c) \log (b-k+5).$$

By (13) and (8), the last inequality results, for every $\pi \in \Pi$, in

$$H(G'_\pi, P') \lesssim \frac{(k-3)b^{\underline{k-3}}+b^{\underline{k-2}}}{b^{k-2}} \log (b-k+5) = \frac{b^{\underline{k-3}}}{b^{k-3}} \log (b-k+5).$$

Comparing this with (10)-(12) yields the desired result. □

5. The method. While the problem treated in the last section seems to be outside the reach of the Fredman-Komlós technique, we should not forget that the present proof of Theorem FK is just a streamlined version of the original argument in [4], put in a broader prospective. In fact, the Fredman-Komlós approach is based on two crucial ideas. First, the sets (or, for the other problem treated in their paper, pairs of sets) to be separated are imbedded as edges into a suitable graph. In the present proof of Theorem FK, this idea, and even the graph G, are as in [4]. Then, in a way, the "volume bound," (cf. [4]), is applied to the edges. More precisely, at this point, in [4], the concept of "strong coloring" of the vertices of a graph is introduced, and an information-theoretic inequality is derived. We would like to explain why this inequality, the Theorem 1 in [4], is, in case of very simple graphs, a combination of our Theorem 1, Lemma K and Corollary 1. In fact, our concept of graph entropy is implicitly rediscovered for those graphs in [4].

Let G be an arbitrary graph and let V be a graph with only two connected components; a complete subgraph and an isolated point. Let V^d be the dth power of V, (cf. Part 1). The mapping $f : \mathbf{V}(G) \to V^d$ is said to be a *strong coloring* of G if the edges of G are preserved by f. Fredman and Komlós's key theorem is, in our language,

LEMMA IT. *If $f : \mathbf{V}(G) \to V^d$ is a strong coloring of the graph G, then*

$$\log \frac{|\mathbf{V}(G)|}{\alpha(G)} \leqq \sum_{i=1}^{d} H(V, P_i)$$

where, for every vertex v in V, $P_i(v)$ is the fraction of those vertices in G for which the ith coordinate of the value of f is v.

Proof. We claim that for an arbitrary G and distribution P on $\mathbf{V}(G)$,

$$H(G, P) \geqq H(P) - \log \alpha(G). \tag{14}$$

In fact, let X, Y achieve $H(G, P)$. By Theorem 1, we have

$$H(G, P) = I(X \wedge Y) = H(X) - \sum_{y \in \mathbf{Y}} H(X \mid Y = y) P_Y(y).$$

Since $H(X \mid Y = y)$ is the entropy of some distribution on the elements of the independent set y of G, it cannot exceed $\log |y|$. Thus,

$$H(G, P) \geqq H(X) - \log \alpha(G) = H(P) - \log \alpha(G).$$

This proves (14).

Let P be the equidistribution on $\mathbf{V}(G)$. Then $H(P) = \log |\mathbf{V}(G)|$, and (14) implies

$$H(G, P) \geqq \log \frac{|\mathbf{V}(G)|}{\alpha(G)}. \tag{15}$$

On the other hand, let V_i^d be a graph on the vertex set of G in which the vertices x and x' of G are adjacent if and only if the ith coordinates of $f(x)$ and $f(x')$ are adjacent in V. Observe that

$$V^d = \bigcup_{i=1}^{d} V_i^d,$$

and, by the definition of graph entropy, we have

$$H(G, P) \leqq H(V^d, P),$$

where P is the distribution generated by f on the vertices of V^d in the natural way. Thus, by Lemma 1, we have

$$H(G, P) \leqq \sum_{i=1}^{d} H(V_i^d, P) \leqq \sum_{i=1}^{d} H(V, P_i),$$

where the last step is an obvious consequence of Lemma 2. □

Not having to rely on graphs of which all the connected components are complete subgraphs allowed us to introduce a more direct approach. The key to this approach had to be, however, a functional defined for arbitrary graphs: graph entropy. To prove a further illustration of this concept, let us dwell a little bit on the problem of the content of a graph, introduced in [4].

Let V be the graph containing a single edge plus an isolated point and let $f: \mathbf{V}(G) \to V^d$ be a strong coloring of G. In [4], the *content* (G, f) is defined to be the fraction of the nonisolated points among all the coordinates of all the sequences in $f(\mathbf{V}(G))$. The *content of the graph* G is defined as the minimum of content (G, f) over all the strong colorings of G which map into powers of this particular V. Fredman and Komlós note that their Theorem 1 gives the estimate

$$\log \frac{|\mathbf{V}(G)|}{\alpha(G)} \leqq \text{content}\,(G).$$

Since we can rely on graph entropy, we will not replace it by its lower bound (15). In fact,

$$\text{content}\,(G) \geqq H(G, P),$$

where P is the equidistribution on $V(G)$.

Finally, let us discuss how tight these bounds are. We shall limit this discussion to Theorem FK. It is easily seen that in our proof of this theorem the upper bound on $H(G_\pi, P)$ is tight in an asymptotic sense for an equipartition f_π. Hence, if we restrict ourselves to equipartitions, our estimate might be asymptotically tight provided that we have equality in

$$H(G, P) \leqq \sum_{\pi \in \Pi} H(G_\pi, P). \tag{16}$$

The question of additivity of graph entropy seems to be quite difficult. A modest attempt to tackle it was made in [6]. There is some reason to hope that for independently chosen equipartitions f_π, graph entropy is additive with large probability, provided that there are not too many of them. However, we have no proof of this. At any rate, we believe to have exhibited graph entropy and in particular, its subadditivity to be at the core of the Fredman–Komlós technique.

We should mention that the two problems treated in [4] have a common generalization to which the same technique applies. Other applications of graph entropy in combinatorics and coding theory will be discussed in a forthcoming paper.

Added December 13, 1985. Using essentially the above approach, the author and Kati Marton have improved upon the Fredman–Komlós bounds. The results will be published in a forthcoming paper in the European Journal of Combinatorics. This and some other recent results we had obtained using this approach have led us to believe that graph entropy yields an efficient method to prove nonexistence results in combinatorics.

Acknowledgments. During the preparation of this paper the author had many invaluable discussions with Kati Marton to whom he would like to express his heartfelt gratitude. Thanks are also due to Andrea Sgarro of Trieste University for encouragement and help and to Gábor Simonyi for having spotted an error in the computations of the proof of Theorem 1.

REFERENCES

[1] C. Berge, *Graphs and Hypergraphs*, 2nd ed., North-Holland, Amsterdam, 1976.

[2] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1982.

[3] P. Erdös and Chao Ko-R. Rado, *Intersection theorems for systems of finite sets*, Quart. J. Math. Oxford Ser. (2), XII (1961), pp. 313-320.

[4] M. Fredman and J. Komlós, *On the Size of Separating Systems and Perfect Hash Functions*, SIAM J. Alg. Disc. Meth., 5 (1984), pp. 61-68.

[5] J. Körner, *Coding of an information source having ambiguous alphabet and the entropy of graphs*, in Trans. 6th Prague Conf. on Inform. Theory, Academia, Prague, 1973, pp. 411-425.

[6] J. Körner and G. Longo, *Two-Step Encoding of Finite Memoryless Sources*, IEEE Trans. Inform. Theory, 19 (1973), pp. 778-782.

[7] R. McEliece, *The Theory of Information and Coding*, Addison-Wesley, Reading, Mass., 1977.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 4, October 1986

007

OPTIMAL NUMBERINGS OF AN $N \times N$ ARRAY*

GRAEME MITCHISON† AND RICHARD DURBIN†

Abstract. Given a numbering of the vertices of a graph, one can define the edgesum [6] as the sum of differences between adjacent vertices. The problem of finding numberings which are optimal in the sense of minimizing the edgesum is NP-complete [2] but has been solved in the special case where the graph is the 2^n cube [3] and for several instances of graphs with high degrees of symmetry [6]. We find the solutions for numberings of an $N \times N$ array. These have practical application in the problem of representing spatial information in a one-dimensional medium. To find our solutions, we exploit the fact that such numberings can always be taken to be ordered, in the sense that numbers increase along rows and down columns. We also consider a generalization of this problem to the case where the differences are raised to a power q. We derive bounds on the edgesum in this case, and show that the optimal numberings for $q < 1$ must be essentially different from those we have found for $q = 1$. While the latter may be assumed to be ordered, and have regions of numbering by rows or columns, neither statement is true for the case $q < 1$. We hypothesize that the solution in this case has a fractal character.

Key words. edgesum, graph numbering, fractal mapping

AMS(MOS) subject classifications. 05B99, 05C99

1. Introduction.

Suppose we number an $N \times N$ array with the integers $1 \cdots N^2$. What numbering minimizes the absolute values of differences between adjacent entries? More precisely, suppose f is a numbering of an $N \times N$ array, with $f_{i,j}$ the number assigned to the (i, j)th entry. Define the "cost" of f, $C(f)$ by $C(f) = \sum \Delta_{ij}$, where $\Delta_{ij} = |f_{i,j+1} - f_{i,j}|^q + |f_{i+1,j} - f_{i,j}|^q$ and $q > 0$. Our goal is to find numberings f which minimize $C(f)$. The significance of the exponent q is that it indicates the extent to which large jumps between adjacent squares are penalized ($q > 1$) or tolerated ($q < 1$). We shall derive the optimal numbering for the case $q = 1$ (Fig. 4), and obtain bounds on $C(f)$ for other values of q. We also consider generalizations to higher dimensions.

This problem arises in many practical contexts, for instance in representing two-dimensional arrays on a sequential file in a computer. Suppose one wishes to perform local calculations around a point in the array, as in the case of evaluating a differential operator. Then Δ_{ij} measures the distance in the file which must be traversed in each local operation. An analogous problem arises in computer hardware, when one wishes to place components of a multi-dimensional array processor on a lower dimensional chassis.

Our original interest in this problem arose with a biological question. The cortex of the brain of higher mammals can be regarded as a sheet of nerve cells. In the part of the cortex devoted to vision, cells respond to certain visual stimuli, such as oriented bars of light against a dark background. A major discovery of recent years is that variables used to describe these stimuli, such as the location of an edge in space, or its orientation, are mapped in a systematic manner on the cortex [7]. This mapping of more than two variables onto a two-dimensional sheet is, in some cases, accomplished by cycling through the values of variables to give striped patterns. This suggests that the nervous system may be trying to achieve as much continuity as possible in mapping these variables onto the cortex. The numbering of an array represents the most simplified mathematical model of this problem.

The higher dimensional generalization is relevant to coding theory. If the numbers $0 \cdots 2^m$ are encoded as strings of m 1's or 0's, the mean absolute change which results

* Received by the editors May 24, 1985 and in revised form January 2, 1986.

† MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England.

from an error in one bit is proportional to $\sum \Delta_{ij}$, $q=1$, the sum being taken over adjacent vertices of the 2^m cube. Harper [3] showed that the numbering of the 2^m cube that corresponds to the standard binary code is optimal, for all m. When $q=2$, $C(f)$ is the mean-square error, and this too was shown to be minimized by the standard binary code [1]. As we shall see, the analogous numbering of N^m (the expression of a number in base N) is *not* optimal in the case $q=1$ when N is large enough.

2. The case $q=1$. It is convenient to describe the $N\times N$ array geometrically as a square. Numberings by rows or columns, such as $f_{ij}=(i-1)N+j$, give $C(f)=N^3-N$. These numberings are ordered, in the sense that $f_{ij}<f_{i,j+1}$ and $f_{ij}<f_{i+1,j}$.

PROPOSITION 1. *For $q\geqq 1$, if f is any numbering, then there is an ordered numbering g with $C(g)\leqq C(f)$.*

Proof. It is clear that the horizontal differences within a row are minimized by ordering that row. It follows that if we order all rows we can only decrease the horizontal contribution to C. To show that horizontal ordering also improves the vertical contributions, we proceed inductively down the rows. Suppose that we have ordered a row and performed the appropriate permutation on the rows below so that every element retains the same vertical neighbour. Suppose i and j lie next to each other in our ordered row, and that i lies above m, j above n, with $i<j$, $m>n$. Then the vertical differences are reduced by switching m and n in the lower row, by virtue of the inequality $|i-n|^q+|j-m|^q\leqq|i-m|^q+|j-n|^q$ which holds for such sets of i, j, m, n when $q\geqq 1$. We can clearly order the row below by a sequence of such permutations.

Call the resulting horizontally ordered numbering h, and order this vertically to obtain g. It remains to show that g is still horizontally ordered. If not, then for some i, j and k with $i<j$ we would have $g_{ik}>g_{jk}$. Assume $g_{jk}=h_{js}$. There are $(k-1)$ h_{j^*}'s with $h_{j^*}<g_{jk}$, and so, including h_{is}, there are at least k h_{i^*}'s with $h_{i^*}<g_{jk}<g_{ik}$, which is a contradiction. □

We may now assume that, with $q\geqq 1$, all contenders for a minimal numbering are ordered. For $q=1$, and for an ordered f, $C(f)=$ (sum of entries for R.H. (right-hand) column + lower row − L.H. (left-hand) column − top row). This dependence upon the boundary values greatly simplifies our task.

Numbering the $N\times N$ square in sequence, let U denote the region which has been filled when either the L.H. edge or top row is complete. For definiteness, let us assume in the following that the L.H. edge is filled first. Let V denote the region which has yet to be numbered just before the top R.H. corner is numbered. Let S denote the sum of edge elements on the boundary of U.

LEMMA 1. *The largest value of S is obtained by successively filling in from either the top row or the L.H. edge the longest columns or rows within U compatible with the ordering rule.*

Proof. We start by placing 1 in the top L.H. corner and putting 2 in an adjacent square. After this, the longest row or column lies next to the pair 1, 2 (see Fig. 1), and can only be two squares long if it is not to violate ordering. Continuing in this way, if at any stage we do not fill a row or column as far as ordering permits, then the remaining sites must be filled later, and the edge entries which are made before this is done will be smaller than if the row or column had been completed at once. So the largest S is obtained by filling rows or columns as far as allowed.

If at some stage we fill a row or column which is not the longest available, then there must be two rows or columns (one of each) of lengths m and n, $m<n$, which have been filled consecutively, with that of length m filled first. Interchanging the order of filling increases S. □

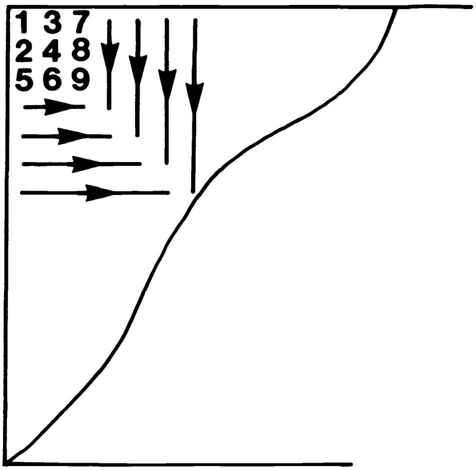


FIG. 1. *The initial sequence for optimally filling region U.*

COROLLARY. *S is maximized by filling first the largest square P which lies within U, and then filling in order of decreasing length the columns or rows in the regions to the right of, and below, P. P is filled as a succession of squares of increasing size: there is a choice of adding a row or a column to each completed square, after which a column or row, respectively, must follow.*

We now know how to number U so as to maximize the sum of entries along its boundary. By symmetry, we can number V so as to minimize the corresponding sum, and in this way we can minimize the total cost $C(f)$ for a given pair of regions U and V. It is also not difficult to see that the region between U and V is best filled with vertical columns in sequence from left to right (as permitted by Proposition 1). In this way the entries along the bottom row which do not lie within V are minimized, and likewise the entries along the top row which do not lie in U are maximized.

The next question is how to choose the shapes of U and V so as to give a global minimum for $C(f)$.

LEMMA 2. *Suppose U, V and the region W between U and V have been numbered optimally, in the manner prescribed above. Let P be the largest square in the top left-hand corner inside U, and Q the largest square in the bottom right-hand corner inside V. Then $C(f)$ can be made smaller by deleting from U the region which lies to the right of P and the corresponding region from V to the left of Q.*

Proof. Let U_1 denote the region of U to the right of P, and U_2 that below P. Let V_1 denote the region in V to the left of Q, and V_2 that above Q.

Let us give names to sums of numbers in various segments of the boundary of the $N \times N$ square. E_1 denotes the sum for the segment of the top row above U_1, E_2 that for the left-hand edge of U_2 and E_3 that for the remaining part of the top row to the right of U. F_1, F_2 and F_3 are defined analogously (see Fig. 2a).

We can write $C(f) = (F_1 + F_2 + F_3) - (E_1 + E_2 + E_3) +$ (terms from boundaries of P and Q). Our plan is to remove V_1 and U_1 and show that $(E_1 + E_2 + E_3)$ is increased. By symmetry it will follow that $(F_1 + F_2 + F_3)$ is decreased, and hence that $C(f)$ is decreased.

First remove V_1 from V. That is to say, fill $W \cup V_1$ with columns, and then fill $V - V_1$ optimally. This will increase E_3, but leave terms on the boundary of Q unchanged, since Q is still the maximal square in V and is filled last.

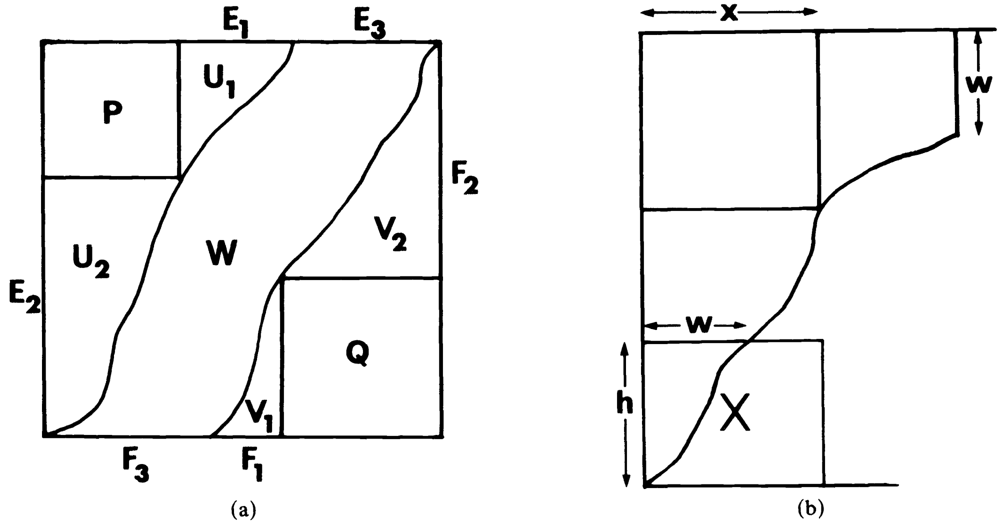


FIG. 2. (a) *The regions defined in Lemma* 2, *and the segments of boundary over which the sums* E_i *and* F_i *are taken.* (b) *Showing that the region* U_i *can be removed column by column.*

Next remove the right-most column of U_1 from U and add it to $W \cup V_1$. Suppose this column has height w. Let the side of P be x. In U_2, suppose the last row with width greater than w lies at height h above the bottom row. Removing the right-most column of U_1 decreases E_2 by hw. By assumption, $W \cup V_1$ is filled with vertical stripes, and this includes the region to the right of U_2 below P, since we have removed V_1 and P and Q are disjoint. The last entry of U_1 in the top row will be increased by at least hx (the area of X in Fig. 2b). In all, $E_1 + E_2$ will be increased by at least $hx - hw > 0$, since $x > w$.

Removing the right-most columns of U_1 step by step, we eventually remove all of U_1 and in doing so increase $(E_1 + E_2)$. This completes the proof. □

LEMMA 3. *Suppose we have put U and V into the above form, and hence that U consists of a square P of side x, no points to the right of this square, and a region* U_2 *beneath P. Then* $C(f)$ *is minimized by making the width of* U_2 *constant down to a height x above the bottom row, and then giving the row at height* $y < x$ *a length of y or* $y - 1$.

Proof. Assume there is a row, r, at a height y above the bottom left-hand corner of length w with $w < y$, $w < x$ (Fig. 3). Since $w < x$, we can assume the row above r is longer than w (if not, take the row above). If we extend r by 1, C decreases by $(y-1) - w > 0$ (if $w = y - 1$ the change is indifferent). Similarly if $w > y$ for some row r, we can assume that the row below is shorter than w, and if we shorten r by 1 then C decreases by $w - y > 0$. □

We now know how to construct the optimal numbering. By the Corollary to Lemma 1, and Lemma 2, the top left-hand corner of U is a square of side x, say, with no points to the right of it. By Lemma 3, this square continues downwards as a rectangle until it is a distance x from the bottom, and then cuts inwards to meet the left-hand column in a 45° triangle. By symmetry, V can be assumed to have the same shape (reflected about the horizontal and vertical midlines), and the region between them is optimally filled with vertical columns in sequence from left to right.

For this numbering we find $C(f) = N^3 - xN^2 + 2x^2N - 2x^2N - 2x^3/3 + O(N^2)$, which is minimized by taking $x = (1 - 2^{-1/2})N$. For this value of x, $C(f) = (4 - 2^{1/2})N^3/3 + O(N^2) = 0.868N^3 + O(N^2)$. Of course, these dimensions are only realized approximately for integral N. The smallest N where the regions U and V exceed

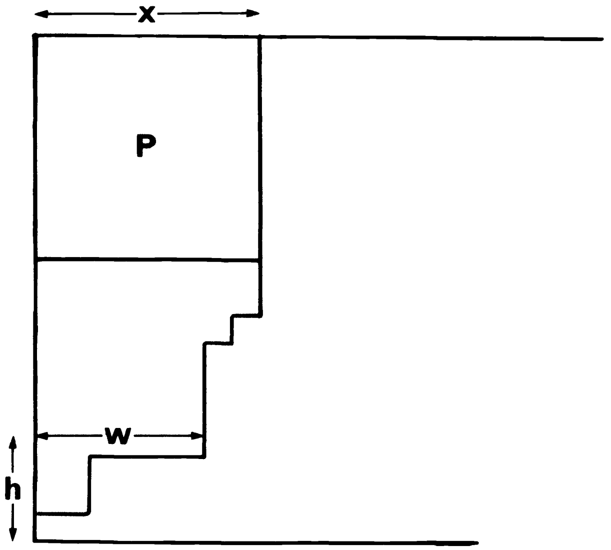


FIG. 3. *Deriving the optimal shape for U.*

a column width and therefore allow a better numbering than by rows or columns is $N=5$ (Fig. 4b).

There are many equivalent variants of the optimal numbering. Proposition 1 shows that ordering a numbering does not increase the cost. Suppose we have two ordered columns (or rows) next to each other. If all the elements of one are larger than those in the other, either column may be reversed without changing the cost, for the differences down each column are unchanged, as is the total difference between the columns. If the two sets of elements overlap in their ordering, reversing one column will increase C, as is easily seen. In the section of columns in the centre of Fig. 4a, any column can be reversed without changing the cost. All other rows or columns overlap in their ordering with their neighbours, so no other reversals are possible.

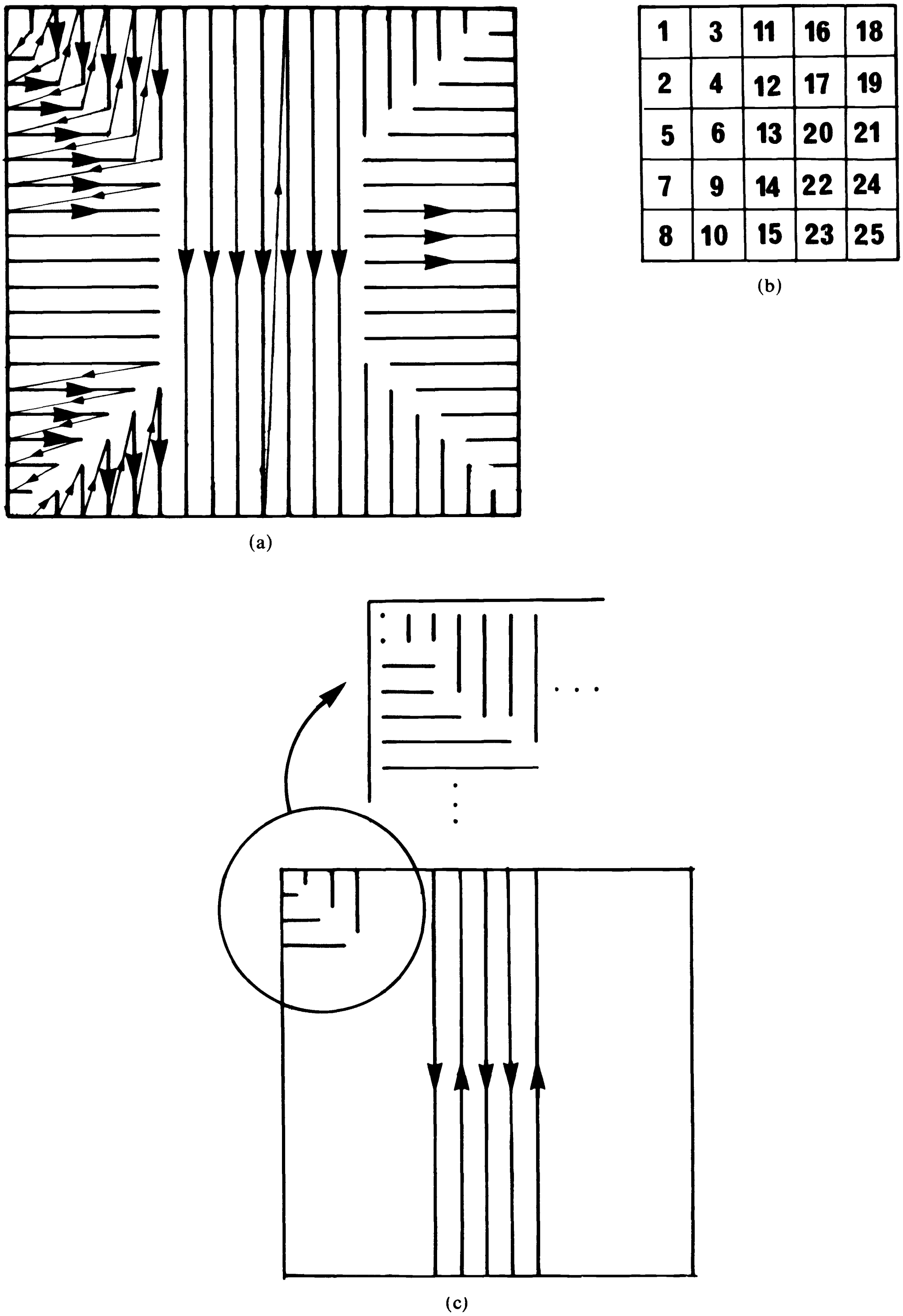
As we have seen in Lemma 3 and the Corollary to Lemma 1 there is a choice along the diagonals in the corners. Each diagonal site may belong to either a row or a column (Fig. 4c). If the numbering is required to be stably symmetric, as defined by Harper [6], then this ambiguity is removed in the top left and bottom right corners, but not in the other two corners.

Finally, we have assumed (in the preamble to Lemma 1) that the L.H. edge is completed before the top row. If this is not so, then the boundary of U includes the top row and that of V the bottom row, the region W between them is best filled with horizontal stripes, and we obtain a construction equivalent to that given above but reflected around the diagonal.

We summarize this section as follows:

THEOREM 1. *For $q=1$, the optimal numberings are given by Fig.* 4a, *its reflected version, and all their variants obtained by reversing the order of the central complete columns and selecting diagonal elements in corners.*

There is no essential difficulty in applying the foregoing arguments to an $N \times N$ torus instead of the square. One can still show that there must exist an optimal numbering which is ordered, in the sense of Proposition 1. This follows from the fact that the minimal cost on a circle is twice the difference between the minimum and the maximum, which is attained by the ordered numbering. Having ordered a circle, the



(a)

1	3	11	16	18
2	4	12	17	19
5	6	13	20	21
7	9	14	22	24
8	10	15	23	25

(b)

(c)

FIG. 4. (a) *Schematic representation of an optimal numbering. The arrows denote the direction in which a column or row is numbered. The light lines show which row follows which column, or vice versa, and otherwise rows are filled downwards or columns to the right.* (b) *An optimal numbering for* $N = 5$. (c) *Variants of the optimal numbering in* (a).

minimum cost is reached by imposing the same ordering on adjacent rows, as in Proposition 1. Now, given an ordered numbering, the cost for a torus is just twice that for the same numbering on a square, and it follows that the numbering derived above is also optimal for a torus. The torus numbering which Harper [5] proposes lacks the diagonal pattern in the bottom left and top right corners and is therefore not optimal. His argument that the regions we call U and V must be rectangles, although intuitively appealing, is in fact unsound.

3. The case $q<1$. We have not been able to prove that any particular numbering is optimal for $q<1$. However, we have been able to obtain lower bounds for $C(f)$, and to find numberings for which C has the same exponent of N as this bound.

Good numberings for $q<1$ are obtained by a different strategy from that used when $q=1$. Instead of trying to spread differences between adjacent squares as evenly as possible, it is better to cluster large differences together, as suggested by the following:

LEMMA 4. *Let s_i, $i=1$ to k, be real numbers satisfying $\sum_1^k s_i = a$, where $a>0$. Then, for $q<1$, $T=\sum_1^k |s_i|^q$ is minimized by taking $s_i=0$ for all i except for some $s_j=a$, giving $T=a^q$. For $q>1$, T is minimized by $s_i=a/k$, giving $T=a^q k^{1-q}$.*

Proof. If any $s_i<0$, distributing it over other positive s_i will decrease T. So we can assume $s_i>0$, all i. The function $\sum |s_i|^q - \lambda \sum s_i$, λ a Lagrange multiplier, has a turning point in $[0, a]^k$ at $s_1=s_2=\cdots=s_k$. One easily checks that this is a minimum only for $q>1$. When $q<1$, the minimum lies on the boundary of $[0, a]^k$. □

We now derive a lower bound for $C(f)$ when $q<1$.

PROPOSITION 2. *When $q<1$, $C(f) \geqq (2q+1)^{-1}.N^{1+2q}+O(N^{2q})$.*

Proof. Join 1 and N^2 by a path constructed by following a horizontal line through 1 until it meets a vertical line through N^2, after which this vertical line is followed (Fig. 5). The interfaces between squares crossed by this path contribute terms to $C(f)$. We can write this contribution as $\sum |s_i|^q$, where the s_i are differences at each interface along the path. By Lemma 4, the minimum contribution of the path between N^2 and 1 is $(N^2-1)^q$. Now look at the region left after the entire horizontal row and vertical column partially used by the path are deleted. The difference between the largest and smallest numbers left on the square must be at least $N^2-2N-2=(N-1)^2-1$.

If we repeat the above construction, the path between these two extremal numbers does not repeat any of the contributions to $C(f)$ made by the previous path (even if they cross, because different interfaces are used by the two paths at their cross-over; see Fig. 5). The second path contributes at least $((N-1)^2-1)^q$. This process may be repeated until the whole square is used up, from which it follows that

$$C(f) \geqq \sum_1^N (i^2-1)^q = (1+2q)^{-1}.N^{1+2q}+O(N^{2q}) \quad \text{as required.} \qquad \square$$

Consider now how a good numbering might be constructed. By Lemma 4, paths give minimal contributions if the difference between their endpoints is concentrated in a single step. We expect to create the smallest set of differences if the steps in a bundle of adjacent paths are lined up. But this then divides the numbers into regions within which the same construction can be repeated. We therefore look for a numbering in which some motif is repeated at a succession of smaller scales—in loose parlance, a "fractal" construction.

Figure 6 shows a simple example of such a construction. Suppose $N=w^{n+1}$. Starting in the top left-hand corner, number sequentially in columns of height w, until a $w \times w$ square is filled. Repeat this w times to make a column of $w \times w$ squares, then replicate this to make w such columns, so filling a square of size $w^2 \times w^2$. The whole $N \times N$ square is filled after $n+1$ iterations of this process.

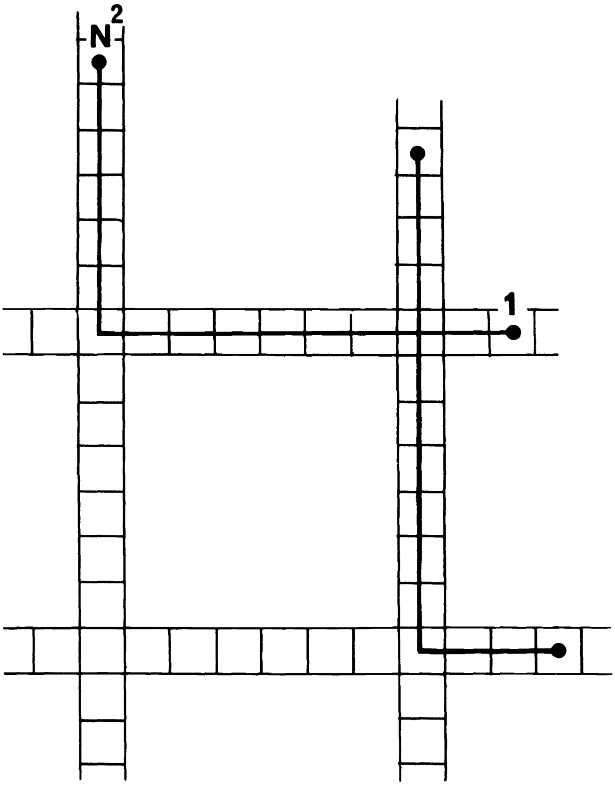


FIG. 5. *Paths used in proof of Proposition* 2.

The cost of this numbering f is easily calculated. Each $w\times w$ square contributes $w(w-1)(w^q+1)$, and there are N^2/w^2 such squares. The boundaries of $w\times w$ squares within a $w^2\times w^2$ square contribute $w^2(w-1)(w^q+1)(w^2-w+1)^q$, and there are N^2/w^4 squares of side w^2. Continuing in this manner, we find

$$\begin{aligned}C(f)/N^2&=(1-1/w)(1+w^q)+(1/w-1/w^2)(1+w^q)(w^2-w+1)^q\\&\quad+(1/w^2-1/w^3)(1+w^q)(w^4-w^3+w^2-w+1)^q+\cdots,\end{aligned}\tag{1}$$

or

$$\begin{aligned}&(C(f)\,.\,(1+w)^q)/(N^2(1+w^q)(1-1/w))\\&\quad=(w+1)^q+(w^3+1)^q/w+(w^5+1)^q/w^2+\cdots+(w^{2n+1}+1)^q/w^n.\end{aligned}\tag{2}$$

Expanding each term on the R.H.S in binomial series gives

$$\begin{aligned}&(C(f)\,.\,(1+w)^q)/N^2(1+w^q)(1-1/w)\\&\quad=w^q\sum_{i=0}^{n}w^{i(2q-1)}+qw^{q-1}\sum_{i=0}^{n}w^{i(2q-3)}+\cdots\\&\quad=w^q(N^{2q-1}-1)/(w^{2q-1}-1)+qw^{q-1}(N^{2q-3}-1)/(w^{2q-3}-1)+\cdots\end{aligned}\tag{3}$$

when $q\neq\frac{1}{2}$.

When $\frac{1}{2}<q<1$ this gives us

$$C(f)=N^{1+2q}.w^q(1+w^q)(1-1/w)/((w^{2q-1}-1)(1+w)^q)+O(N^2).\tag{4}$$

The term in w is minimized by taking $w=2$. When $q=\frac{3}{4}$, for instance, this value gives $C(f)=2.39N^{5/2}+O(N^2)$.

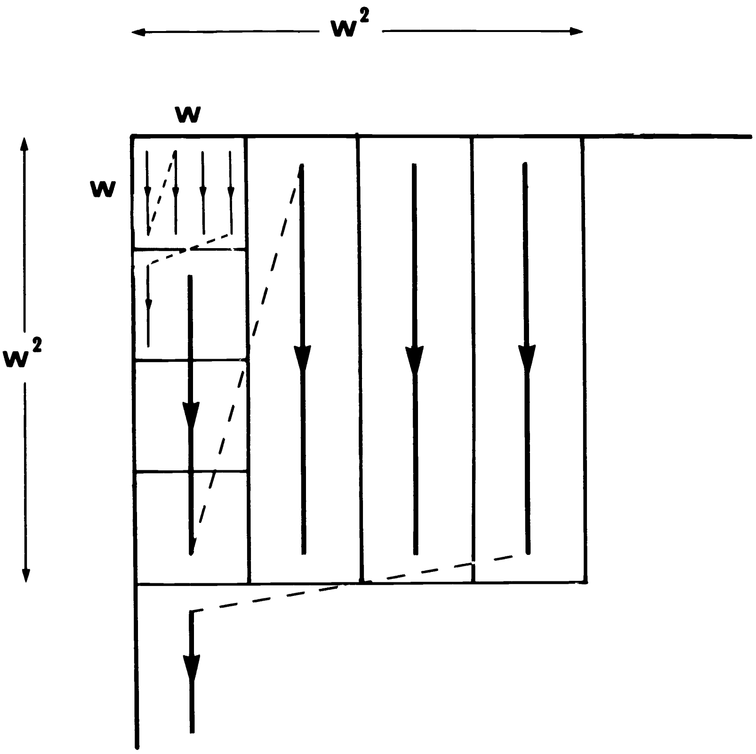


FIG. 6. *A construction for a numbering with cost* $C(f)=O(N^{1+2q})$.

When $q=\frac{1}{2}$ the first series on the R.H.S of (3) reduces to $(n+1)$, and, since $N=w^{n+1}$, we have

$$(5) \qquad C(f)=N^2 \log_w N.(1+w^{1/2})(w^{1/2}-w^{-1/2})/(1+w)^{1/2}+O(N^2).$$

Again, this is minimized by taking $w=2$, which gives $C(f)=.99N^2 \log_2 N+O(N^2)$.

When $q<\frac{1}{2}$, the term in N^{1+2q} no longer dominates, and may be neglected. We have only to count up the terms which contribute to the coefficient of N^2, and the R.H.S. of (2) converges for large n to a sum which depends only upon w and q. For $q=\frac{1}{4}$, for instance, calculations show that C is minimized by taking $w=2$, which gives $C(f)=3.53N^2$. When $q=.1$, C is minimized by $w=4$, with $C(f)=2.39N^2$.

It is striking that, for $q<\frac{1}{2}$, we can find a numbering with a cost proportional to N^2. This says that the average Δ_{ij} does not depend on N. In fact, the coefficients of N^2 which our numbering give are not too far from the best possible. The smallest differences between a given square and its four neighbours are, of course, 1, 2, 3, and 4, and these give a lower bound for the cost of $\frac{1}{2}(1^q+2^q+3^q+4^q)N^2$. Taking $q=\frac{1}{4}$, say, this cost is $2.4N^2$, whereas our numbering gives $3.5N^2$.

What we have shown is that, for $\frac{1}{2}<q<1$, there is a numbering which depends upon N in the same manner as the bound given by Proposition 3. When $q<\frac{1}{2}$, the dependence on N^2 is the best possible. We summarize this in the following:

PROPOSITION 3. *When* $\frac{1}{2}<q<1$, *the minimal cost* C *is* $O(N^{1+2q})$. *When* $q<\frac{1}{2}$, $C=O(N^2)$.

We do not know what are the optimal numberings for $q<1$. The numbering in Fig. 6 is ordered, but Proposition 1 no longer holds when $q<1$, and it seems likely that the best numbering is unordered. A simple example is the square $\begin{smallmatrix}1&2\\3&4\end{smallmatrix}$ with $q=\frac{1}{2}$. This has a cost $2+2.2^{1/2}$, while the unordered square $\begin{smallmatrix}1&2\\4&3\end{smallmatrix}$, which is optimal, has cost $3+3^{1/2}$.

Note that the optimal numberings for $q<1$ cannot contain a rectangle of stripes with sides a constant factor of N, since this would contribute a term of order N^{2+q}. In fact, there cannot be a block of stripes with sides of order $N^{(1+2q)/(2+q)}$. In this sense, the optimal numberings must have fine-scale detail.

4. The case $q>1$. The optimal numbering for $q=1$ (Fig. 4a) is no longer optimal for $q>1$. This numbering generates terms of $O(N^{1+2q})$, which arise from boundaries of the central region of stripes and the diagonals of the bottom left and top right corners. A row or column numbering gives $C(f)=N^{1+q}(N-1)+N(N-1)=N^{2+q}+O(N^{1+q})$, with a lower exponent for N when $q>1$. In fact, this is the best possible exponent for N.

PROPOSITION 4. *For* $q>1$, $C(f)\geqslant 4.2^{-q}.N^{2+q}/(q+2)$.

Proof. By Proposition 1, we may assume that f is ordered. Join the points 1 and N^2 by an elbow-shaped path (Fig. 7). As f is ordered, these points lie in diagonally opposite corners, so the path has length $2N-2$. By Lemma 4, this path contributes $(N^2-1)^q.(2N-2)^{1-q}$ to $C(f)$. Now consider the remaining square after removing the path. The smallest and largest entries in this square have a difference of at least $(N-1)^2-1$ and can be connected by a similar elbow-shaped path of length $2N-4$. Continuing this construction, we get a total contribution to $C(f)$ from these paths of at least

$$\sum_{i=1}^{N}(i^2-1)^q.(2i-2)^{1-q}\sim 2^{1-q}\sum_{i=1}^{N} i^{q+1}\sim 2^{1-q}.N^{2+q}/(q+2).$$

But there is another set of paths at right angles to these (dotted in Fig. 7) which shares

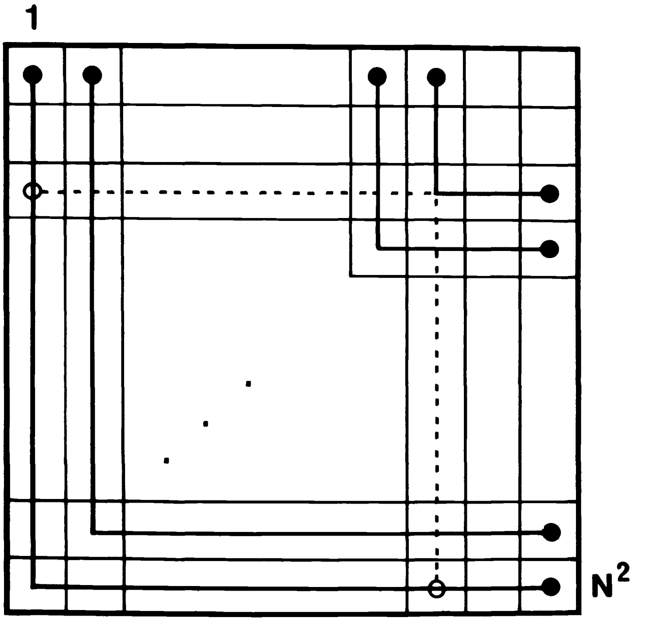


FIG. 7. *Paths used in proof of Proposition* 4.

no interfaces, and we can put the same lower bound on its contribution to $C(f)$. The result follows from this □.

Are the row or column numberings optimal for $q > 1$? This is not the case for $q > 2$, for the diagonal numbering, $f_{ij} = i + (i+j-1)(i+j-2)/2$, gives $C(f) \sim 4N^{2+q}/(q+2)$, which is less than N^{2+q} when $q > 2$. Moghadam [9] has shown that, for rectangles of all dimensions, the diagonal mapping is optimal for $q = \infty$, i.e., for the problem of minimizing the maximum difference between adjacent vertices.

5. Generalizations to higher dimensions. Let $\mathbf{i} = (i_1, \cdots, i_s)$ denote an element in the s-dimensional array N^s, and let f be a numbering of N^s. We can define $C(f) = \sum \Delta_{\mathbf{i}}$ where $\Delta_{\mathbf{i}}$ is the sum of differences $|f_{\mathbf{i}} - f_{\mathbf{j}}|^q$ over the s positive neighbours $\mathbf{j}$ of $\mathbf{i}$. What numberings minimize $C(f)$?

Consider first the case $q = 1$. When $N = 2$, Harper [3] showed that the "natural numbering", which assigns to the vertex $(i_1, \cdots, i_s)$ the binary number $i_1 + 2i_2 + \cdots + 2^{s-1}i_s$, is optimal, for all s. For arbitrary N, the analogue of this is $f(\mathbf{i}) = i_1 + Ni_2 + \cdots + N^{s-1}i_{s-1}$, which is just a generalized row or column numbering (by 0 to $N^s - 1$). We know this is not best when $s = 2$, $N > 5$ (Fig. 4b). It follows that the generalized row or column numberings are not best for any s when $N > 5$. For let g be our optimal numbering for $s = 2$ (Fig. 4a). Then we can define a numbering for any $s > 2$ dimensions by $f(\mathbf{i}) = f(i_1, \cdots, i_s) = i_1 + Ni_2 + \cdots + N^{s-2} . g(i_{s-1}, i_s)$, which is clearly better than the generalized row numbering.

How small a cost can be achieved for the s-cube, N^s, when $q = 1$? Can one do better, when $s > 2$, than the minimum for $s = 2$? The answer to the latter question is yes. Given a numbering n_s for N^s with cost c_sN^{2s-1} and $a < N/2$, construct a numbering n_{s+1} for N^{s+1} as follows:

$$n_{s+1}(i_1, \cdots, i_{s+1}) = \begin{cases} an_s(i_1, \cdots, i_s) + i_{s+1}, & i_{s+1} < a, \\ n_s(i_1, \cdots, i_s) + N^s i_{s+1}, & a < i_{s+1} < N - a, \\ N^s(N-a) + an_s(i_1, \cdots, i_s) + i_{s+1}, & N - a < i_{s+1}. \end{cases}$$

This has cost $c_{s+1}N^{2s+1} = N^s(N^s(N-a) + a) + 2a^2c_sN^{2s+1} + (N-2a)c_sN^{2s-1}$. Hence $c_{s+1} = 1 - a/N + 2c_sa^2/N^2 + O(N^{-1})$, which has a minimum of $(1 - 1/(8c_s))$ when $a/N = 1/(4c_s)$. So c_s rapidly converges to $(2 + 2^{1/2})/4$, which is slightly smaller than $(4 - 2^{1/2})/3$, which was the minimum we obtained for $s = 2$.

Considerable progress has been made for the case $q = \infty$, sometimes called the bandwidth problem. For general graphs this is known to be NP-complete [10]. As mentioned in the previous section, Moghadam [9] has shown that the optimal numbering for rectangles is the diagonal numbering, extending a result of Harper [4] for the 2^s cube.

Another type of generalization involves replacing the four nearest neighbours in the sum for $C(f)$ by a larger set of (possibly weighted) neighbours. Lindsey [8] has shown that when the set of "neighbours" of a point i includes all points which agree with i in all but one coordinate, then a row or column numbering is best for N^s.

Finally, it is interesting to ask whether the composition of two optimal maps is itself optimal, as in the case of the map $f: N^4 \to (N^2)^2$ defined by $f(i, j, k, l) = (g(i, j), g(k, l))$, where g is an optimal numbering of $N \times N$.

Acknowledgments. We thank Professor S. W. Golomb for useful comments and for broadening our cultural horizons, and Professor L. H. Harper for many valuable suggestions.

REFERENCES

[1] T. R. Crimmins, H. M. Horowitz, C. J. Palermo and R. V. Palermo, *Minimization of mean-square error for data transmitted in group codes*, IEEE Trans. Inform. Theory, IT-15, 1 (1969), pp. 72-78.

[2] M. R. Garey, D. S. Johnson and L. Stockmeyer, *Some simplified NP-complete graph problems*, Theoret. Comput. Sci., 1 (1976), pp. 237-267.

[3] L. H. Harper, *Optimal assignments of numbers to vertices*, J. Soc. Indust. Appl. Math., 12 (1964), pp. 131-135.

[4] ———, *Optimal numberings and isoperimetric problems on graphs*, J. Combin. Theory, 1 (1966), pp. 385-393.

[5] ———, *A necessary condition on minimal cube numberings*, J. Appl. Probab., 4 (1967), pp. 397-401.

[6] ———, *Stabilization and the edgesum problem*, Ars Combin., 4 (1977), pp. 225-270.

[7] D. H. Hubel and T. N. Wiesel, *Brain mechanisms of vision*, Sci. American, 241(3) (1979), pp. 150-162.

[8] J. H. Lindsey, *Assignment of numbers to vertices*, Amer. Math. Monthly, 71 (1964), pp. 508-516.

[9] H. Moghadam, *Compression operators and a solution to the bandwidth problem of the product of n paths*, Thesis, Univ. of Calif. at Riverside, Riverside, CA, 1983.

[10] C. H. Papadimitriou, *The NP-completeness of the bandwidth minimization problem*, Computing, 16 (1976), pp. 263-270.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 4, October 1986

008

AN APPROXIMATION TO THE STATIONARY DISTRIBUTION OF A NEARLY COMPLETELY DECOMPOSABLE MARKOV CHAIN AND ITS ERROR BOUND*

MOSHE HAVIV† AND Y. RITOV†

Abstract. In Haviv (Ph.D. dissertation, Yale Univ., New Haven, CT, 1983) an approximation procedure for computing the stationary distribution of a nearly completely decomposable (NCD) Markov chain is suggested. There and in Haviv (this Journal, 7 (1986), pp. 589-593) the incurred error is analyzed. In particular, a series expansion for the error is developed. Courtois and Semal (J. Assoc. Comput. Mach., 31 (1984), pp. 804-825) independently of us, replaced this point approximation with a set of points. Using algebraic methods, they proved that the exact distribution lies in the convex set spanned by this set. We give a probabilistic interpretation for this set and then obtain their results in a more elementary way. We compute the convex combination leading to the exact distribution and develop a bound on it. Finally, we show how approximation to this convex combination leads to an error reduction in a current approximation. It is the first time that a probabilistic approach is made in order to analyze NCD Markov chains.

AMS(MOS) subject classifications. 60J10, 15A51

1. Introduction and summary. A Markov chain is called nearly completely decomposable (NCD) if its state space can be partitioned into a number of subsets in a way that transitions are most likely to occur between states in a same subset, while transitions between states belonging to different subsets are much rarer. The special structure of NCD Markov chains can be stated in terms of the stochastic matrices which describe them. We call such a matrix an NCD stochastic matrix. Up to a permutation of rows and corresponding columns, an NCD stochastic matrix is characterized by diagonal blocks which are "almost" stochastic and off-diagonal blocks having relatively smaller entries. In particular, by permuting rows and corresponding columns, we may assume that P is given in the following form:

$$P=\begin{pmatrix} P_{J(1)} & P_{J(1)J(2)} & \cdots & P_{J(1)J(q)} \\ P_{J(2)J(1)} & P_{J(2)} & \cdots & P_{J(2)J(q)} \\ \vdots & \vdots & & \vdots \\ P_{J(q)J(1)} & P_{J(q)J(2)} & \cdots & P_{J(q)} \end{pmatrix} \tag{1.1}$$

which represents a partition $\underset{\sim}{J}=[J(1), J(2), \cdots, J(q)]$, of the state space to q subsets.[1]

NCD models are common. They usually exist in computer systems. There, transition rates are fast in the "inner" side and slow in the "outer" side. Hence, groups of states that can reach each other rapidly will form a subset. Such a partition will result in an NCD representation of the system. For some examples, the reader is referred to Courtois [2], Zarling [16], W. J. Stewart [13], Vantilborgh [15] and the references cited therein.

* Received by the editors March 18, 1985, and in revised form January 9, 1986.

† Department of Statistics, Hebrew University, Jerusalem, Israel.

[1] Note that for $J, K \subseteq \underset{\sim}{J}$, P_{JK} is the submatrix of P with rows index by J and columns indexed by K. Also we use P_J for P_{JJ}.

In the sequel we will assume that the Markov chain represented by P has only one recurrent class. This implies that P has a unique stationary vector satisfying $yP=y$, $yu=1$ and $y \geqq 0$ (where u denotes a column vector all its entries are equal to 1). The vector y represents the long run behavior of the chain. Throughout we also assume that each subset $J \in \underset{\sim}{J}$ contains at least on recurrent state. This is equivalent to assuming that for each $J \in \underset{\sim}{J}$, $y_i > 0$, for some state $i \in J$.

We next introduce some notation that applies to the matrix P. Fix subset $J \in \underset{\sim}{J}$. The subvector of y corresponding to subset J is denoted y_J. Since $y_J \neq 0$, we can normalize it so that the sum of its components is one. This normalized vector is called the conditional stationary distribution (CSD) of subset J and will be denoted z_J. This vector can be viewed as the asymptotic probabilistic behavior within a subset. Also the stationary probability of subset J is the quantity $\sum_{i \in J} y_i$, which will be denoted k_J. The vector $k=(k_{J(1)}, \cdots, k_{J(q)})$ can be viewed as the asymptotic probability distribution of being in various subsets. Of course,

$$y_J = k_J z_J.$$

Finally, we need the following notation. For a subset J, let J' be its complement. Hence, the matrix P can be represented (perhaps after permuting rows and corresponding columns) by

$$P = \begin{pmatrix} P_J & P_{JJ'} \\ P_{J'J} & P_{J'} \end{pmatrix}.$$

The NCD structure for Markov chains frequently occurs when the state space is very large. Hence, one looks for approximations procedures for approximating y while reducing the computational burden. Usually, the approximations are based on two steps. The first approximates z_J for all $J \in \underset{\sim}{J}$, while the second approximates k. This approach was originally taken by Simon and Ando [12] and further developed by Courtois [1], Vantilborgh [15], G. W. Stewart [14], Courtois and Semal [3], Haviv and Van der Heyden [7] and others. They all start by constructing matrices whose dominant eigenvectors are (small) perturbations of z_J, $J \in \underset{\sim}{J}$ or of k. A different approach, based on probabilistic arguments for approximating z_J was suggested by Haviv [6]. Next we state a slight modification to that procedure.

Fix a subset $J \in \underset{\sim}{J}$ and a row vector π in the $(|J|)$-dimensional simplex. Let $A \equiv (I-P_J)^{-1}$. The inverse of $I-P_J$ exists and equals $\sum_{m=0}^{\infty} P_J^m$ (Kemeny and Snell [5, p. 46]). Hence, $(\pi A)_j$ for $j \in J$, stands for the expected number of time epochs the system visits state j before it first leaves subset J, given that it initiates at subset J with π as the marginal distribution over its states. Similarly, $\pi A u$ denotes the expected number of time epochs for visiting subset J before first leaving it under the same initial conditions. Thus, $[z_J(\pi)]_j \equiv (\pi A)_j / \pi A u$ is the ratio between the number of visits to state j and the number of visits to subset J, both before first leaving subset J and under the same initial conditions represented by π. In the NCD case, one expects $[z_J(\pi)]_j$ to be a reasonable approximation to $[z_J]_j$. The rationale is as follows. $[z_J]_j$ gives the asymptotic fraction of time periods the process visits state $j \in J$ out of the number of time periods it visited subset J. In the NCD case, the horizon until the first exit from the initial subset is fairly long to scramble the arbitrary initial conditions.

Before proceeding we would like to note here that the matrix A can be replaced with the adjoint matrix of $I-P_J$, $\operatorname{adj}(I-P_J)$. This is true since $A = \operatorname{adj}(I-P_J)/\det(I-P_J)$. This fact not only reduces the computational burden, but also shows that almost singularity of $I-P_J$, (i.e., $\det(I-P_J)$ being close to zero) does not cause numerical difficulties.

By converting the model to algebraic terms, Courtois and Semal [3] have shown that z_J lies in the convex hull spanned by $\{z_J(e_i), 1 \leqq i \leqq |J|\}$, where e_i is the ith unit vector. Hence, they have induced that $\text{Min}_i\, A_{ij}/a_i \leqq [z_j]_j \leqq \max_i A_{ij}/a_i$, where $a \equiv Au$. We next obtain these results using elementary probabilistic arguments. More specially, we show that $z_J(\pi)$ is the conditional stationary distribution of subset J in the $(|J|+1)$-dimensional Markov chain represented by the transition matrix

$$P_J(\pi) \equiv \begin{pmatrix} P_J & P_{JJ'}u \\ \pi & 0 \end{pmatrix}.$$

Then we show that there exists a choice of π, say π^* such that $z_J(\pi^*) = z_J$, which immediately leads to the above bounds. Moreover, we show that a choice of π^* is the normalization of $y_{J'}P_{J'J}$.

Our final result concerns showing how a natural choice of π as a function of current approximation to y lead to a new approximation to z_J which results in a serious error reduction. Moreover, we bound the reduction from above.

Finally, we would like to refer the reader to Haviv [6, § 2.4] and Haviv [8]. There a series expansion for $z_J(\pi) - z_J$ is developed.

2. The approximation and its error bounds.

THEOREM 1. *Let the* $(|J|+1) \times (|J|+1)$ *transition matrix* $P_J(\pi)$ *be as defined above. Then* $z_J(\pi)$ *is the conditional stationary distribution of subset J with respect to* $P_J(\pi)$.

Proof. Denote by s the state appended to subset J to construct the Markov chain represented by $P_J(\pi)$. To simplify notation, denote $P_J(\pi)$ by Q and let p be its stationary distribution. Finally, denote by $_sQ^n_{sj}$ the probability that the Markov process represented by the transition matrix Q while initiating at state s visits state j at the nth epoch without passing through state s again previously.

Bearing in mind the definition of A, we have $(\pi A)_j = \sum_{n=0}^{\infty} {_sQ^n_{sj}}$. The last equals $\lim_{m\to\infty} (\sum_{n=0}^{m} Q^n_{sj} / \sum_{n=0}^{m} Q^n_{ss})$ (Karlin and Taylor [4, p. 35]). Hence,

$$\begin{aligned}(\pi A)_j/\pi a &= \lim_{m\to\infty} \left(\sum_{n=0}^{m} Q^n_{sj} \Big/ \sum_{k\in J} \sum_{n=0}^{m} Q^n_{sk} \right) \\ &= \lim_{m\to\infty} \frac{1}{m} \sum_{n=0}^{m} Q^n_{sj} \Big/ \lim_{m\to\infty} \frac{1}{m} \sum_{k\in J} \sum_{n=0}^{m} Q^n_{sk} \\ &= p_j \Big/ \sum_{k\in J} p_k\end{aligned}$$

which by definition equals $[z_J(\pi)]_j$, completing the proof. □

The heuristic behind the theorem is as follows. Instead of looking at the expected number of visits at a state j before leaving the subset J, we look at the long run average of the number of visits at state j in a process which is equivalent to the original process except that any time the chain leaves J it returns to it through the vector of marginal probabilities π.

Remark 1. We would like to note that if one replaces the zero in the southeast corner of $P_J(\pi)$ by some α, $\alpha \in [0, 1)$, and accordingly replaces π with $(1-\alpha)\pi$, the corresponding conditional stationary distribution of subset J will be preserved. We omit a formal proof of this observation.

Remark 2. By the fact that the conditional stationary distribution of subset J in $P_J(\pi)$ is the stationary distribution of $P_J + P_{JJ'}u\pi$ (Kemeny and Snell [5, p. 115]), one can see that Theorem 1 and the conclusion of Courtois and Semal [3, p. 817] are equivalent.

THEOREM 2. *There exists a π, say π^*, such that $z_J(\pi^*)=z_J$; moreover, a choice for π^* is*

$$\pi^* = \frac{y_{J'}P_{J'J}}{y_{J'}P_{J'J}u}.$$

Proof. Since $y=yP$ or

$$(y_J, y_{J'}) = (y_J, y_{J'})\begin{pmatrix} P_J & P_{JJ'} \\ P_{J'J} & P_{J'} \end{pmatrix},$$

one easily gets that $y_J = y_J P_J + y_{J'}P_{J'J}$ and that $y_{J'} = y_J P_{JJ'} + y_{J'}P_{J'}$. By postmultiplying both sides of the last equality with u, one gets that $k_{J'} = y_J P_{JJ'}u + y_{J'}P_{J'}u$. Hence

$$(y_J, k_{J'}) = (y_J, k_{J'})\begin{pmatrix} P_J & P_{JJ'}u \\ \dfrac{y_{J'}P_{J'J}}{k_{J'}} & \dfrac{y_J{}'P_{J'}u}{k_{J'}} \end{pmatrix}.$$

To the stochastic matrix in the last equality corresponds a matrix of the type $P_J(\pi)$ (see Remark 1) with π being the normalization of $y_{J'}P_{J'J}$ which is, of course, $y_{J'}P_{J'J}/y_{J'}P_{J'J}u$. □

Remark 3. The above choice for π^* is just the stationary marginal probabilities of entering subject J through any of its states from subset J', i.e.,

$$\pi_j^* = \lim_{n\to\infty} P\{S_n = j \mid S_{n-1}\in J', S_n \in J\}$$

when S_n denotes the state of the system at stage n.

Remark 4. By combining Remark 2 and Theorem 2, one can easily see that z_J is the stationary distribution of $P_J + P_{JJ'}u\pi^*$. This observation is an alternative proof to the fact that "exact aggregation" exists. Also note by Hunter [9, Cor. 4.1.2], that for any rank one matrix $A = wv$ where w and v are column and row vectors, respectively, such that $vw=1$, with $y_J w \neq 0$ and $vu \neq 0$ and such that $I - P_J - A$ is stochastic with stationary distribution y_J, satisfies $y_J = v(I-P_J)^{-1}/v(I-P_J)^{-1}u$.

Remark 5. Theorems for a different choice for π^* can be deduced from Courtois and Semal [3, p. 809].

LEMMA. *For any π in the $(|J|)$-dimensional simplex, there exists a vector δ is the same simplex which satisfies*

$$(\pi A)_j/\pi a = \sum_i \delta_i A_{ij}/a_i$$

for all $j\in J$.

Proof. For a given π set $\delta_i = \pi_i a_i/\pi a$, for all $i \in J$. □

THEOREM 3. *For any π, $z_J(\pi)$ lies in the convex hull spanned by $\{z_J(e_i), i\in J\}$. In particular, for any norm $\|\cdot\|$, $\|z_J - z_J(\pi)\| = \|z_J(\pi^*) - z_J(\pi)\| \leqq \mathrm{Max}_{i\in J} \|z_J(e_i) - z_J(\pi)\|$. Hence, for all $j\in J$*

$$\underset{i}{\mathrm{Min}}\, A_{ij}/a_i \leqq [z_J(\pi)]_j \leqq \underset{i}{\mathrm{Max}}\, A_{ij}/a_i.$$

Proof. The theorem follows immediately from Theorem 2 and the lemma. □

COROLLARY. $\mathrm{Min}_i\, A_{ij}/a_i \leqq (z_J)_j \leqq \mathrm{Max}_i\, A_{ij}/a_i$.

Proof. The corollary follows immediately from Theorems 2 and 3. □

Remark 6. Theorem 3 and the corollary can be deduced also from Theorem 6 of Courtois and Semal [3, p. 810]. They also showed that $\mathrm{Max}_i A_{ij}/a_i = A_{jj}/a_j$. This intuitive fact can be proved also in an elementary probability way. The heuristic argument is that when we count the number of time epochs the system stays in set J, we can count separately the number of time epochs until the first visit in j and from that event on. Making the first period shorter should maximize the fraction of time epochs the system visits state j until the first time it leaves subset J. We omit the details.

For any π let $\delta(\pi)$ be the vector one gets through the transformation indicated in the lemma. Hence $z_J(\pi) = \delta(\pi)Y$, where the matrix Y is defined by $Y_{ij} = A_{ij}/a_i$. To simplify notation, we use δ instead of $\delta(\pi)$. In particular, $\delta^* = \delta(\pi^*)$. The following theorem bounds the difference between $z_J(\pi) - z_J$ in terms of $\delta - \delta^*$ and the matrix Y.

THEOREM 4. *Let $\gamma = \delta - \delta^*$; hence $\gamma u = 0$. Also, for any norm $\|\cdot\|_\alpha$ on $R^{|J|}$, let the operator $\tau_\alpha(\cdot)$ be defined by*

$$\tau_\alpha(Y) = \text{Max } \|xY\|_\alpha,$$
$$\text{s.t. } xu = 0,$$
$$\|x\| \leqq 1.$$

Then

$$\|z_J(\pi) - z_J\| \leqq \|\gamma\|_\alpha \tau_\alpha(Y).$$

In particular,

$$\tau_{l_1}(Y) = \tfrac{1}{2} \underset{i,k}{\text{Max}} \sum_j |Y_{ij} - Y_{kj}|$$

and

$$\tau_{l_\infty}(Y) = \underset{j}{\text{Max}} \sum_i |Y_{ij} - Y_{(m_j)j}|$$

where $Y_{(m_j)j}$ is a median of the entries in the jth column of the matrix Y.

Proof. Since $z_J(\pi) - z_J = \delta Y - \delta Y^* = \gamma Y$ and since $\gamma u = 0$, the theorem's result for an arbitrary norm follows. The explicit forms for the l_1 and the l_∞ norm appear in Rothblum [11]. □

We would like to note that the bound of Theorem 4 is tight. It is easy to see from the definition of $\tau_\alpha(\cdot)$ that a vector δ such that the inequality is replaced by equality exists.

Let $\bar{y}$ be approximation to y, let $\bar{\pi} = \bar{y}_{J'} P_{J'J} / \bar{y}_{J'} P_{J'J} u$ and let γ be $\delta(\bar{\pi}) - \delta(\pi^*)$, namely γ is the difference between the $\delta(\pi)$ corresponding to $\bar{y}$ and $\delta(\pi^*)$, the $\delta(\pi)$ corresponding to the exact y. Of course, $\|\gamma\|_\alpha$ is a measure representing the accuracy of $\bar{y}$ as an approximation to y. Then the lemma shows that while premultiplying Y by $\delta(\pi)$ in order to get $z_J(\pi)$, the error shrinks by a factor of $\tau_\alpha(Y)$.

The above discussion shows that the smaller $\tau_\alpha(Y)$ is the larger the error reduction is. Since by assumption P represents an NCD Markov chain, P can be written in the form $P = P^* + \varepsilon C$, where P^* is completely decomposable and ε is a "small" number. The approximation for the conditional stationary distributions discusssed here are of the order of $O(\varepsilon)$ (Haviv, [8]), namely they approach the stationary distribution of P^* as P approaches P^* linearly in the direction C. This, of course, will be true for e_i, $1 \leqq i \leqq |J|$ as a choice for π, namely for any Y_{ij} as an approximation for $(z_J)_j$. Hence, terms in the same columns of Y will differ by the order of $O(\varepsilon)$. In this case, $\tau_{l_1}(Y)$ and $\tau_{l_\infty}(Y)$ will be of the order of $O(\varepsilon)$ as well.

REFERENCES

[1] P. J. COURTOIS, *Error analysis in nearly-completely decomposable stochastic systems*, Econometrica, 43 (1975), pp. 691-709.

[2] P. J. Courtois, *Decomposability*, Academic Press, New York, 1977.

[3] P. J. Courtois and P. Semal, *Bounds for the positive eigenvectors of nonnegative matrices and for their approximations by decomposition*, J. Assoc. Comput. Mach., 31 (1984), pp. 804-825.

[4] S. Karlin and H. M. Taylor, *A Second Course in Stochastic Processes*, Academic Press, New York, 1981.

[5] J. C. Kemeny and J. L. Snell, *Finite Markov Chains*, D. Van Nostrand Company, New York, 1960.

[6] M. Haviv, *Approximations in Markov chains and Markov decision models*, Ph.D. dissertation, Yale University, New Haven, CT, 1983.

[7] M. Haviv and L. Van der Heyden, *Perturbation bounds for the stationary probabilities of a finite Markov chain*, Adv. in Appl. Probab., 16 (1984), pp. 804-818.

[8] M. Haviv, *An approximation to the stationary distribution of a nearly completely decomposable Markov chain and its error analysis*, this Journal, 7 (1986), pp. 589-593.

[9] J. J. Hunter, *Generalized inverses and their application to applied probability problems*, Linear Algebra Appl., 45 (1982), pp. 159-198.

[10] D. F. McAllister, G. W. Stewart and W. J. Stewart, *A two-stage iteration for solving nearly uncoupled chains*, Technical Report 1984, Department of Computer Science, Univ. of Maryland, College Park, MD, 1983.

[11] U. G. Rothblum, *Explicit solution to optimization problems of the intersection of the unit ball of the l_1 and the l_∞ norms with hyperplane*, this Journal, 5 (1984), pp. 619-632.

[12] H. A. Simon and A. Ando, *Aggregation of variables in dynamic systems*, Econometrica, 29 (1961), pp. 111-138.

[13] W. J. Stewart, *A comparison of numerical techniques in Markov modeling*, Comm. ACM, 2 (1978), pp. 144-152.

[14] G. W. Stewart, *Computable error bounds for aggregated Markov chains*, J. Assoc. Comput. Mach., 30 (1983), pp. 271-285.

[15] H. Vantilborgh, *Aggregation with an error of $O(\varepsilon^2)$*, J. Assoc. Comput. Mach., 32 (1985), pp. 161-190.

[16] R. L. Zarling, *Numerical solutions of nearly decomposable queuing networks*, Ph.D. dissertation, Dept. of Computer Sciences, University of North Carolina, Chapel Hill, NC, 1976.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 4, October 1986

009

AN APPROXIMATION TO THE STATIONARY DISTRIBUTION OF A NEARLY COMPLETELY DECOMPOSABLE MARKOV CHAIN AND ITS ERROR ANALYSIS*

MOSHE HAVIV†

Abstract. All existing approximations for the stationary distribution of a nearly completely decomposable Markov chain are based on solving systems which are perturbations of the exact system. We develop here an original approximation which is based on probabilistic intuition. A series expansion of the accrued error is given as well.

AMS(MOS) subject classifications. 60J10, 15A51, 15A09

1. Introduction. A Markov chain is called nearly completely decomposable (NCD) if its state space can be partitioned into a number of subsets in a way that transitions are most likely to occur between states in a same subset, while transitions between states belonging to different subsets are much rarer. The special structure of NCD Markov chains can be stated in terms of the stochastic matrices which describe them. We call such a matrix an NCD stochastic matrix. Up to a permutation of rows and corresponding columns, an NCD stochastic matrix is characterized by diagonal blocks which are "almost" stochastic and off-diagonal blocks having relatively smaller entries. In particular, by permuting rows and corresponding columns, we may assume that P is given in the following form:

$$(1.1)\qquad P=\begin{pmatrix} P_{J(1)} & P_{J(1)J(2)} & \cdots & P_{J(1)J(q)} \\ P_{J(2)J(1)} & P_{J(2)} & \cdots & P_{J(2)J(q)} \\ \vdots & & & \vdots \\ P_{J(q)J(1)} & P_{J(q)J(2)} & \cdots & P_{J(q)} \end{pmatrix}$$

which represents some partition $\underset{\sim}{J}=\{J(1), J(2),\cdots,J(q)\}$, of the state space to q subsets.[1]

In the sequel, we will be more specific about the size of the elements in these matrices, though a precise definition will not be needed. For more details on nearly complete decomposability, see Simon and Ando [14] and Courtois [3].

ASSUMPTION A. The Markov chain represented by P has only one recurrent class.

Assumption A implies that P has a unique stationary vector y satisfying

$$(1.2)\qquad yP=y,\quad yu=1,\quad y\geqq 0$$

(where u denotes the vector all of whose entries equal to 1). Of course, the vector y is the asymptotic probability distribution of the corresponding Markov chain. Throughout we also assume the following.

ASSUMPTION B. Each subset $J\in\underset{\sim}{J}$ contains at least one recurrent state.

Assumption B is equivalent to assuming that for each $J\in\underset{\sim}{J}$, $y_i>0$, for some state $i\in J$.

We next introduce some notation that applies to the matrix P. Fix subset $J\in\underset{\sim}{J}$. The subvector of y corresponding to subset J is denoted y_J. Since $y_J\neq 0$, we can

* Received by the editors February 12, 1985, and in final revised form January 9, 1986.

† Department of Statistics, Hebrew University, Jerusalem, Israel.

[1] Note that for $J, K\subseteq\underset{\sim}{J}$, P_{JK} is the submatrix of P with rows indexed by J and columns indexed by K. Also we use P_J for P_{JJ}.

normalize it so that the sum of its components is one. This normalized vector is called the conditional stationary distribution (CSD) of subset J and will be denoted z_J. This vector can be viewed as the asymptotic probabilistic behavior within a subset. Also, the stationary probability of subset J is the quantity $\sum_{i\in J} y_i$ which will be denoted k_J. The vector $k=(k_{J(1)},\cdots,k_{J(q)})$ can be viewed as the asymptotic probability distribution of being in various subsets. Of course,

$$y_J = k_J z_J. \tag{1.3}$$

The NCD structure for Markov chains frequently occurs when the state space is very large. Hence, one looks for approximation procedures for approximating y which reduce the computational burden. Usually the approximations are based on two steps. The first approximates z_J for all $J\in \underset{\sim}{J}$ while the second approximates k. See e.g., Simon and Ando [14], Courtois [2], Vantilborgh [16], Stewart [15] and Haviv and Van der Heyden [6]. They all are based on constructing systems of linear equations which are (small) perturbations of the systems whose solutions are z_J, $J\in\underset{\sim}{J}$ and k. For details see Haviv [5].

Next we define an original approximation procedure for approximating z_J which is based on probabilistic arguments.[2]

Fix a subset $J\in\underset{\sim}{J}$ and let $A\equiv(I-P_J)^{-1}$. The inverse of $I-P_J$ exists and equals $\sum_{m=0}^{\infty} P_J^m$ (Kemeny and Snell [8, p. 46]). In particular, A_{ij} is the expected number of visits in state j before first leaving subset J, given that the chain initiates at state $i\in J$. Accordingly, $a_i\equiv\sum_j A_{ij}$ is the expected number of transitions in subset J before subset J is left, given that the chain starts in state $i\in J$. Thus, A_{ij}/a_i can be interpreted as the ratio between the expected number of visits in state $j\in J$ and the expected number of visits in subset J, both before the first exit out of the subset and given that the chain starts in state $i\in J$. In the NCD case, one expects this approximation to $(z_J)_j$ to be reasonable since $(z_J)_j$ gives the asymptotic fraction of time periods the process visits state $j\in J$ out of the number of time periods it visits subset J. In the NCD case, the horizon until the first exit from the initial subset is fairly long to scramble the initial condition. Of course, different initial states i might yield (slightly) different approximations. Also $\sum_i \pi_i A_{ij}/a_i$ for any π such that $\pi^t u=1$, is a reasonable approximation to $(z_J)_j$.

We like to note here that the matrix A can be replaced with the adjoint matrix of $I-P_J$, adj $(I-P_J)$. This is true since $B^{-1}=\text{adj}\,(\text{B})/\det\,(\text{B})$. This observation not only reduces the computational burden, but also shows that the almost singularity of $I-P_J$, i.e., $\det(I-P_J)$ being close to zero, does not cause any numerical difficulties.

In the next section we look at some properties of the suggested approximation. In particular, we give a series expansion of this approximation as a function of the probability of leaving subset J and conclude that the incurred error is of the order of this probability.

Recently we have learned of a paper by Courtois and Semal [4]. They proved there that z_J is contained in the convex hall spanned by the set of approximation attained by the choices of π, $\{\pi=e_i, 1\leqq i\leqq |J|\}$, where e_i is the ith unit vector, hence, having bounds for the error incurred by the approximation. Independently, in Haviv and Ritov [7] by using probabilistic arguments, we obtained the same results and further developed them to get

$$\text{Min}_i\, A_{ij}/a_i \leqq (z_J)_j \leqq A_{jj}/a_j.$$

2. The approximation and its error analysis. Before proceeding we need the following lemma.

[2] This paper is based on Haviv [5, § 2.4].

LEMMA 1 [8, p. 115]. *For a subset $J \in \underline{J}$ and its complement J' express P (perhaps after permuting columns and corresponding rows) by*

$$P = \begin{pmatrix} P_J & P_{JJ'} \\ P_{J'J} & P_{J'} \end{pmatrix}.$$

Then the $|J| \times |J|$ matrix $\tilde{P}_J \equiv P_J + P_{JJ'}(I - P_{J'})^{-1} P_{J'J}$ is stochastic with stationary distribution z_J.

Since $(I - P_{J'})^{-1}$ is nonnegative, one can see by the lemma that $\tilde{P}_J = P_J + \tilde{T}$ for some $\tilde{T} \geqq 0$. In the NCD case, $\tilde{T}$ has small terms. To emphasize this fact, we let $\tilde{T} = \varepsilon T$ for a small ε and a nonnegative matrix T. Also let $A(\varepsilon) \equiv (I - \tilde{P}_J + \varepsilon T)^{-1}$ and let $a(\varepsilon) \equiv A(\varepsilon)u$. Hence, the suggested approximation for $(z_J)_j$ is $A_{ij}(\varepsilon)/a_i(\varepsilon)$ for some arbitrary i. We next obtain, under some technical conditions, a series expansion for $A_{ij}(\varepsilon)/a_i(\varepsilon)$ around $\varepsilon = 0$. In particular, the leading coefficient of that series expansion will be $(z_J)_j$. Thus, the suggested approximation is an $O(\varepsilon)$ one. Thus, one can easily get from that a series expansion for the incurred error, namely for $A_{ij}(\varepsilon)/a_i(\varepsilon) - (z_J)_j$.

The following notation follows Rothblum [11]. Let $S \equiv I - \tilde{P}_J$. Then, zero is an eigenvalue of S. Moreover, as $\tilde{P}_J$ is assumed to be irreducible, the multiplicity of this eigenvalue of S is one. The unique (up to a multiplicative constant) corresponding left and right eigenvectors are z_J and u, respectively. The index of a matrix B is defined as the smallest integer t such that the null spaces of B^{t+1} and B^t coincide. It is known that the index of S is 1 (cf. Campbell and Meyer, [1, p. 152]). The projection on the space spanned by the eigenvectors of a matrix B belonging to the eigenvalue zero is denoted by E and is called the eigenprojection at zero. Finally, for a matrix B, the Drazin inverse is defined as $(B - E)^{-1}(I - E)$.

We next state the promised series expansion for $A_{ij}(\varepsilon)/a_i(\varepsilon)$.

THEOREM 1. *Suppose the matrix T is nonsingular and that the index of $T^{-1}S$ is one.*[3] *Let D be the Drazin inverse of $T^{-1}S$. Then for ε small enough,*

$$\frac{A_{ij}(\varepsilon)}{a_i(\varepsilon)} = \frac{(z_J)_j + z_J T u \left\{ \sum_{k=1}^{\infty} [D^k(-\varepsilon)^k] T^{-1} \right\}_{ij}}{1 + z_J T u \sum_m \left\{ \sum_{k=1}^{\infty} [D^k(-\varepsilon)^k] T^{-1} \right\}_{im}}$$

$$= (z_J)_j - z_J T u \left[(DT^{-1})_{ij} - (z_J)_j \sum_m (DT^{-1})_{im} \right] \varepsilon + O(\varepsilon^2).$$

Proof. As z_J and u are, respectively, left and right eigenvectors of S belonging to zero, $z_J T$ and u are, respectively, left and right eigenvectors of $T^{-1}S$ belonging to zero. Hence, the assumption that the index of $T^{-1}S$ is one implies that the eigenprojection of $T^{-1}S$ at zero is $u z_J T / z_J T u$ (Rothblum, [10]). Then Rothblum [11] establishes that for ε small enough,

$$(T^{-1}S + \varepsilon I)^{-1} = u z_J T / z_J T u \varepsilon - \sum_{k=0}^{\infty} D^{k+1}(-\varepsilon)^k$$

and so

$$A(\varepsilon) = (S + \varepsilon T)^{-1} = (T^{-1}S + \varepsilon I)^{-1} T^{-1} = u z_J / z_J T u \varepsilon - \sum_{k=0}^{\infty} [D^{k+1}(-\varepsilon)^k] T^{-1}$$

and the theorem follows by straightforward division. □

[3] The condition that the index of $T^{-1}S$ is one is satisfied in the cases where zero is a simple eigenvalue of $T^{-1}S$ (cf. Campbell and Meyer, [1, p. 133]). Since zero is a simple eigenvalue of S it is most likely that the same will be true for $T^{-1}S$.

As Theorem 1 shows, for any choice of i, $A_{ij}(\varepsilon)/a_i(\varepsilon)$ is an $O(\varepsilon)$ approximation to $(z_J)_j$. Evidently, for any vector π with $\pi^t u = 1$, $\sum_i \pi_i A_{ij}(\varepsilon)/a_i(\varepsilon)$ is also an $O(\varepsilon)$ approximation to $(z_J)_j$. It is easy to see that a choice of π such that $\sum_i \pi_i [DT^{-1})_{ij} - (z_J)_j \sum_m (DT^{-1})_{im}] = O(\varepsilon)$ yields an $O(\varepsilon^2)$ approximation. Next we show that if π is chosen to be an $O(\varepsilon)$ approximation to $z_J T$, then the corresponding error is $O(\varepsilon^2)$.

THEOREM 2. *Let $\pi \in R^{|J|}$ be an $O(\varepsilon)$ approximation to $z_J T$. Then, under the same technical assumptions given in Theorem* 1, *$\sum_{i\in J} \pi_i A_{ij}(\varepsilon)/a_i(\varepsilon)$ is an $O(\varepsilon^2)$ approximation to $(z_J)_j$.*

Proof. As indicated in the proof of Theorem 1, $z_J T$ is a left eigenvector of $T^{-1}S$ belonging to the eigenvalue zero. Hence, it is also a left eigenvalue of its Drazin inverse D (cf., [1]), namely, $(z_J T)D = 0$. Thus, if $\pi = z_J T + O(\varepsilon)$ then $\pi D T^{-1} = O(\varepsilon)$ and hence $\pi(DT^{-1})_j - (z_J)_j \pi(DT^{-1})u = O(\varepsilon)$ where the latter times $z_J T u$ agrees with the coefficient of ε in the series expansion of $\sum_i \pi_i A_{ij}(\varepsilon)/a_i(\varepsilon)$. □

Theorem 2 above has only a theoretical merit since the matrix T, or an $O(\varepsilon)$ approximation to it, are not available while constructing the approximation. In particular, as shown in Lemma 1, the computation of T needs information not contained in P_J. Of course, an $O(\varepsilon^2)$ approximation to $(z_J)_j$ which is based only on P_J does not exist: P_J itself is only an $O(\varepsilon)$ approximation to $\tilde{P}_J$, namely to the transition probabilities leading to z_J. In Vantilborgh [16] one can find an $O(\varepsilon^2)$ approximation to z_J and indeed he uses additional information from outside subsets. In Courtois and Semal [4] and in Haviv and Ritov [7] one can find different choices for π leading to the exact z_J.

We would like to conclude this paper by showing how to develop a series expansion to the corresponding approximation of k, the stationary subset probabilities. First we need the following lemma and notation.

LEMMA 2 (trivial). *Let the $q \times q$ stochastic matrix Q be defined by*

$$Q_{ij} = \sum_{s\in J(i)} (z_{J(i)})_s \sum_{t\in J(j)} P_{st}, \qquad 1 \leqq i, j \leqq q.$$

Then k is the stationary distribution of Q.

The matrix Q, of course, can be approximated by $Q(\varepsilon)$ where $Q(\varepsilon)$ is constructed like Q is but where the $z_{J(i)}$'s $1 \leqq i \leqq q$, are replaced by some $O(\varepsilon)$ approximations to them like those suggested in this paper, which we generally denote by $z_{J(i)}(\varepsilon)$. In other words,

$$Q_{ij}(\varepsilon) = z_{J(i)}(\varepsilon) P_{J(i)J(j)} u, \qquad i \leqq i, j \leqq q.$$

Of course, a given series expansion for the $z_{J(i)}$'s immediately yields a series expansion for $E(\varepsilon) \equiv Q(\varepsilon) - Q$ with $E(\varepsilon) = \sum_{s=1}^{\infty} \varepsilon^s A_s$ for some matrices A_s, $s = 1, 2, \cdots$. Also, let $k(\varepsilon)$ be the stationary distribution of $Q(\varepsilon)$ and let Y be the Drazin inverse of Q. Then, by Schweitzer's [12] perturbation result,[4] for ε small enough,

$$\begin{aligned} k(\varepsilon) &= k \sum_{m=0}^{\infty} (E(\varepsilon) Y)^m \\ &= k \sum_{m=0}^{\infty} \left(\sum_{s=0}^{\infty} \varepsilon^s A_s Y \right)^m \\ &= k \sum_{s=0}^{\infty} \varepsilon^s B_s \end{aligned}$$

[4] Schweitzer's results in terms of the Drazin inverse are given in Meyer [9].

for $B_0 = I$ and some matrices B_s, $s = 1, 2, \cdots$ where B_s is a function of A_t, $t = 1, 2, \cdots, s$ and of Y. Now a series expansion of, say, $y_J(\varepsilon) \equiv k_J(\varepsilon) z_J(\varepsilon)$ can be easily derived. For a different way to obtain series expansions for $y_J(\varepsilon)$, see Courtois [2] and Schweitzer [13].

Acknowledgment. The author would like to acknowledge helpful comments of an anonymous referee.

REFERENCES

[1] S. L. CAMPBELL AND C. D. MEYER, JR., *Generalized Inverses of Linear Transformations*, Pitman Publishing Limited, London, 1979.

[2] P. J. COURTOIS, *Error analysis in nearly-completely decomposable stochastic systems*, Econometrica, 43 (1975), pp. 691-709.

[3] ———, *Decomposability*, Academic Press, New York, 1977.

[4] P. J. COURTOIS AND P. SEMAL, *Bounds for the positive eigenvectors of nonnegative matrices and their approximations by decomposition*, J. Assoc. Comput. Mach., 31 (1984), pp. 804-825.

[5] M. HAVIV, *Approximations in Markov chains and Markov decision models*, Ph.D. dissertation, Yale University, New Haven, CT, 1983.

[6] M. HAVIV AND L. VAN DER HEYDEN, *Perturbation bounds for the stationary distribution of a finite Markov chain*, Adv. in Appl. Probab., 16 (1984), pp. 804-818.

[7] M. HAVIV AND Y. RITOV, *An approximation to the stationary distributions of a nearly completely decomposable Markov chain and its error bound*, this Journal, 7 (1986), pp. 583-588.

[8] J. G. KEMENY AND J. L. SNELL, *Finite Markov Chains*, D. Van Nostrand Company, New York, 1960.

[9] C. D. MEYER, *The role of the group generalized inverses in the theory of Markov chains*, SIAM Rev., 17 (1975), p. 443.

[10] U. G. ROTHBLUM, *Computation of the eigenprojection of a nonnegative matrix at its spectral radius*, in Mathematical Programming Studies 6, Stochastic Systems: Modeling, Identification and Optimization II, Roger J.-B. Wets, ed., 1976, pp. 188-201.

[11] ——— *Resolvant expansions of matrices and applications*, Linear Algebra Appl., 38 (1981), pp. 33-49.

[12] P. J. SCHWEITZER, *Perturbation and finite Markov chains*, J. Appl. Probab., 5 (1968), pp. 401-413.

[13] ———, *Perturbation series expansions of nearly completely decomposable Markov chains*, Working paper series no 8122, Graduate School of Management, University of Rochester, NY, 1981.

[14] H. A. SIMON AND A. ANDO, *Aggregation of variables in dynamic systems*, Econometrica, 29 (1961), pp. 111-138.

[15] G. W. STEWART, *Computable error bounds for aggregated Markov chains*, J. Assoc. Comput. Mach., 30 (1983), pp. 271-285.

[16] H. VANTILBORGH, *Aggregation with an error of $O(\varepsilon^2)$*, J. Assoc. Comput. Mach., 32 (1985), pp. 161-190.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 4, October 1986

010

COMPUTING THE STRUCTURAL INDEX*

I. S. DUFF† AND C. W. GEAR‡

Abstract. The index of many differential/algebraic equations (DAEs) is determined by the structure of the system, that is, by the pattern of nonzero entries in the Jacobians. This paper considers an important subclass of DAEs which can be solved by backward differentiation methods if their index does not exceed two. For this reason, it is desirable to determine whether the index exceeds two or not. In this paper we present an algorithm that determines if the index is one, two, or greater, based only on the structure. The algorithm can be exponential in its execution time: we do not know whether it is possible to get an asymptotically faster algorithm. However, in many practical problems, this algorithm will execute in polynomial time.

Key words. sparse matrices, transversals, differential/algebraic equations, matrix index, index of nilpotency

1. Introduction. For a more detailed discussion of the index of differential/algebraic equations, the reader is referred to [1]-[4]. Here, we summarize the meaning of the index and treat the simple case

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}, t) + \mathbf{G}\mathbf{z}, \tag{1.1a}$$

$$\mathbf{H}\mathbf{y} = \mathbf{A}\mathbf{z}. \tag{1.1b}$$

The dimension of $\mathbf{y}$ and $\mathbf{f}$ is n; the dimension of $\mathbf{z}$ is m. The constant matrices $\mathbf{A}, \mathbf{G}$, and $\mathbf{H}$ have dimensions m by m, n by m, and m by n, respectively.

The index of a differential/algebraic equation system is defined as the minimum number of times the system must be differentiated with respect to the independent variable so that the first derivatives of the all dependent variables can be determined uniquely at a point when only the values of the dependent variables (but not their higher derivatives) are known at that point. Thus, if m is zero in (1.1), $\mathbf{z}$ does not appear and $\mathbf{y}'$ is determined uniquely by (1.1a). In this case, the index is zero. However, if m is nonzero, $\mathbf{z}'$ is not determined by (1.1) so the index is at least one. If we differentiate (1.1b) once, we find that

$$\mathbf{A}\mathbf{z}' = \mathbf{H}\mathbf{y}' = \mathbf{H}\mathbf{f}(\mathbf{y}, t) + \mathbf{H}\mathbf{G}\mathbf{z}. \tag{1.2}$$

If $\mathbf{A}$ is nonsingular, (1.2) determines $\mathbf{z}'$. Hence, a necessary and sufficient condition for the index to be one is that $\mathbf{A}$ is nonsingular. A necessary and sufficient condition for the index to be less than three is that

$$\operatorname{rank}\begin{pmatrix} \mathbf{A} \\ \mathbf{NHG} \end{pmatrix} = m, \tag{1.3}$$

where $\mathbf{N}$ is a full rank r by $m (r \leqq m)$ matrix whose r rows span the left null space of $\mathbf{A}$, that is, $\mathbf{NA} = 0$ and $\mathbf{N}$ is the largest rank matrix with this property. Condition (1.3) follows by differentiating (1.2) once, premultiplying by $\mathbf{N}$, and substituting (1.1a) to get

$$\mathbf{NHG}\mathbf{z}' = -\mathbf{NH}\left[\frac{\partial \mathbf{f}}{\partial \mathbf{y}}(\mathbf{f} + \mathbf{G}\mathbf{z}) + \frac{\partial \mathbf{f}}{\partial \mathbf{t}}\right] \tag{1.4}$$

and solving for $\mathbf{z}'$ from (1.2) and (1.4).

* Received by the editors March 28, 1985, and in revised form February 18, 1986. This work was supported in part by U.S. Department of Energy grant DEAC0276ERO2383 and in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under contract W-31-109-Eng-38. This report has also been published as AERE Harwell Report CSS 173.

† Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, on leave from AERE Harwell, Didcot, Oxon, England.

‡ Department of Computer Science, University of Illinois, Urbana, Illinois 61801.

Clearly the index depends on the actual numerical values of the matrix entries. In many cases, the determination of the index given the numerical values of the matrix is a poorly conditioned problem since it involves rank determination. However, for almost all values of the nonzero entries of the matrices, the index has the same value and is determined by the nonzero structure of the matrices. We call this the *structural index*; its determination is a well-conditioned problem. Since the index of many practical problems assumes its structural value, the structural index is a useful quantity to calculate in a differential/algebraic equation code.

Before examining the structural index further, it is helpful to consider the *structural rank* of a sparse matrix. This is the numerical rank of the matrix for almost all values of its nonzero entries, and is equal to the length of a *maximum transversal.* A transversal is a selection of nonzero elements such that no more than one element is selected from any row or column. A maximum transversal is a transversal of maximum length. It follows that a matrix **A** of order m is nonsingular if and only if a maximum transversal has length m (this is called a *full transversal*). An algorithm for finding a maximum transversal is given in [5]. Hence the problem of determining if the index is one has already been solved, and the algorithm in [5] for doing this has worst case complexity $O(m\tau)$, where τ is the number of nonzeros in the matrix. Clearly, adding additional nonzero entries to a sparse matrix will not reduce its structural rank but may increase it, and the structural rank of a dense matrix is the minimum dimension of the matrix.

The concept of a structural index is more complex than that of structural rank. If **A** is nonnull and dense, the structural index of (1.1) is one. However, it does not follow that adding additional nonzeros to a sparse **A** cannot increase the structural index: it may decrease, increase, or leave the structural index unchanged. Hence we cannot equate the structural index with the maximum or minimum index over all values of the nonzero elements. We illustrate this with the following example in which the addition of a nonzero element to **A** will increase the structural index. If the matrix **A** has the structure

$$\begin{pmatrix} x & 0 & 0 \\ x & 0 & 0 \\ x & 0 & 0 \end{pmatrix}$$

its left null space **N** has rank 2 and the structural index of (1.1) is at least one. Note that we *cannot* give a unique sparse structure for **N** in general, all that we know in this case is that **N** consists of two linearly independent rows each of which is orthogonal to the first column of **A**. A possible structure for **N** is

$$\begin{pmatrix} x & x & 0 \\ 0 & x & x \end{pmatrix}$$

but there are five other rank two structures possible for **N** with two zeros, and seven structures with fewer zeros. Suppose the matrix product **HG** has the structure

$$\begin{pmatrix} 0 & 0 & x \\ 0 & 0 & 0 \\ 0 & x & 0 \end{pmatrix}$$

then **NHG** can have the form

$$\begin{pmatrix} 0 & 0 & x \\ 0 & x & 0 \end{pmatrix}$$

so (1.3) holds with $m = 3$ and the structural index is two. However, if we change **A** by adding a nonzero in the (2, 1) position to get

$$\begin{pmatrix} x & x & 0 \\ x & 0 & 0 \\ x & 0 & 0 \end{pmatrix},$$

$\mathbf{N}$ now has rank one and takes the form

$$(0 \quad x \quad x)$$

implying that

$$\text{rank}\begin{pmatrix} \mathbf{A} \\ \mathbf{NHG} \end{pmatrix} \leqq 2$$

so that the structural index exceeds two.

Henceforth, we will use the words "linear independence", "nonsingular", etc. to mean structural linear independence, structurally nonsingular, etc., in the sense that there exist values of the nonzero entries of the matrices such that these properties hold. If there exist any values of the matrix entries for which properties of this type hold, they hold for almost all values.

In this paper we will extend the algorithm given in [5] to the determination of whether the index is two. Specifically, we will first solve the problem of determining whether

$$\text{rank}\begin{pmatrix} \mathbf{A} \\ \mathbf{NB} \end{pmatrix} = m, \tag{1.5}$$

where $\mathbf{B}$ is an m by m matrix whose structure is also given and $\mathbf{N}$ spans the left null space of $\mathbf{A}$. Then we will extend the algorithm to (1.3). In § 2 we will present the modified algorithm, and in § 3 we will prove that the algorithm does determine whether the index is two.

We normally envisage the algorithm being applied to differential/algebraic equations in the form (1.1). However, it can be applied to equations in a more general form directly. Consider

$$\mathbf{Ax}' + \mathbf{HGx} = \mathbf{0}. \tag{1.6}$$

If $\mathbf{A}$ is nonsingular, the index of this problem is clearly zero since (1.6) can be solved for $\mathbf{x}'$ directly. If $\mathbf{A}$ is singular but (1.3) holds, then the index is one. In fact, the index of the system (1.6) is zero or one as the index of system (1.1) is one or two, so that the algorithm described in this paper can be used to determine whether the index of (1.6) does not exceed one. The index of system (1.6) is just the *index of nilpotency* of the matrix pencil $\mathbf{HG} + \lambda\mathbf{A}$ (see [6]).

The structure of null space bases for sparse matrices is investigated at some length in [7] but the concern there is solely with the sparsity of the basis so that any basis is feasible in that case. The main form of basis that we use in our proofs has a triangular submatrix and is termed a myopic null basis in [7]. In our approach, however, the choice of myopic null basis determines the rows of $\mathbf{B}$ ($\mathbf{HG}$) that are considered. It is this added complexity that makes our algorithm exponential.

We end the introduction with a description of the algorithm given in [5] for determining a maximum transversal (and hence the structural rank) because it will be extended in § 2 to compute the index. That algorithm is equivalent to Algorithm I outlined[1] below.

[1] The outline has a breadth-first search in Insert_Row. In [5], a depth-first search is used. The latter seems preferable for implementation; here, however, we are interested in the form of the algorithm.

ALGORITHM I. In the following, T is a (partial) transversal. It consists of a set of pairs (r_i, c_i), $i=1, \cdots, k$, where r_i and c_i are row and column numbers with no row or column repeated, and the matrix entry $a_{r_i c_i}$ is nonzero. Capital letters are used for sets, and lower-case letters refer to integers. The set union operation is indicated with the "+" character, and set formation with braces. Hence, { } is the empty set. The notation $R(T)[C(T)]$ means the set of rows [columns] represented in T, while $R(i)$-$[C(i)]$ is the set of row [column] indices corresponding to the nonzeros in column [row] i. The algorithm is described recursively to prepare for its extension to the more general problem, although it is just as easily expressed using a loop in this case. The data structure T is global; everything else is local. In the algorithm presented below, comments are enclosed in brackets.

The algorithm proceeds a row at a time. At any time, T contains a maximum transversal of the rows examined so far. As each new row is considered, the algorithm looks for a nonzero column that is not yet in T. If the new row has a nonzero in a column not yet in T, that element can be added directly to T. If not, it is necessary to see if a different maximum transversal of the previous rows would permit the addition of an element from the new row to the transversal. Fortunately, this does not require that all transversals be checked; rather, the algorithm considers replacing an element of the current transversal with an element from the new row to see if the row thus removed from the transversal would have an element in a column not yet in the transversal or if a similar replacement could be made for an element in the latter row. This is done in the function Insert_Row.

```
program Compute_Transversal
        [Main program. It initializes $T$ and then calls the subprogram
        Extend_Transversal which cycles through the rows.]
        $T \leftarrow \{\}$
        Extend_Transversal(1)
        [Rank of A is number of pairs in T]
        endprogram Compute_Transversal
subprogram Extend_Transversal($q$)
        [This subprogram extends the current transversal in $T$ to include rows
        $q$ to $m$ of A.]
        if $q > m$ then return endif
        Insert_Row($q$, $Q$)
        Extend_Transversal($q+1$)
        [This recursive call of Extend_Transversal serves to count through the
        rows of A.]
        endsubprogram Extend_Transversal
boolean function Insert_Row($q$, $Q$)
[This function attempts to add an element from row $q$ to the transversal held in
$T$. The set $S$ contains the columns containing nonzeros from row $q$ and from each
row containing a transversal element which we are considering replacing with an
element from another row. Insert_Row is defined as a Boolean function in prepar-
ation for the algorithm in § 2. Similarly, the parameter $Q$ is present because of
its use in the next algorithm. Here we use only the changes produced in $T$]
$S \leftarrow C(q)$
repeat
  if there exists $c$ in $S$ but not in $C(T)$
    then
```

```
            Add row $q$ and column $c$ to $T$
                [This will require some reorganization of $T$ if the $(q, c)$ element is zero.
                At this stage we are guaranteed that: (i) there is a transversal of length
                one larger than that now in $T$ and (ii) that the longer transversal contains
                row $q$ and column $c$. The depth-first search given in [5] identifies the
                needed reorganization more easily.]
            return true
        else
            $Q \leftarrow$ all $r$ such that $(r, c) \in T$ for some $c \in S$
            [$Q$ is the set of rows which have transversal elements which could be
            involved in a replacement.]
            $S \leftarrow S + \{C(j)$ for all $j$ in $Q\}$
        endif
    until $S$ does not increase in size
    [Note: If $S$ did not increase in size, we did not insert an entry of the transversal
    in row $q$, and $Q+\{q\}$ is a set of linearly dependent rows of A. If any step does
    not insert an entry, the matrix is singular.]
    return false
    endfunction Insert_Row
```

2. Determination of the structural index. In this section we describe an algorithm for determining the structural index. The algorithm is an extension of Algorithm I. Each time that Algorithm I finds a linearly dependent set of rows, that set is rank deficient by one only, so it has identified exactly one row of the null space, **N**, of **A**. The column indices corresponding to nonzeros in this row of **N** represent the rows of **B** that could be included in a row of **NB**. They also represent a set of rows of **A**, one of which can be ignored without decreasing the rank of **A**. Thus, it is sufficient to consider a matrix with one of these rows discarded and a row from **NB** appended. The algorithm discards each of the rows in a dependent set in turn and then searches for the next dependent set[2]. In considering the rows of **NB**, we will show in § 3 that it is sufficient to consider replacing the discarded rows of **A** by the corresponding row of **B**.

ALGORITHM II. Algorithm II uses the same notation and data structures as Algorithm I, plus a global Boolean array of size m called Mark. It is used to mark rows of **A** that have been switched with rows of **B**. The algorithm refers to rows "being marked." The function Insert_Row is the same as described in Algorithm I. Note that the parameter Q is an "output" parameter (**var** in Pascal).

```
    program Determine_Index
            $T \leftarrow \{\}$
            if Extend_Transversal(1)
                then if no marks set
                    then index = 1
                    else index = 2
                    endif
                else index > 2
                endif
```

[2] Since there can be as many as $m-1$ dependent sets and the algorithm may check all combinations, it can take exponential time. However, if there are few sets and each set is small, speed is not a problem. In particular, if the sets consist of single entries because rows of A are empty, the time remains polynomial.

```
        endprogram Determine_Index
Boolean function Extend_Transversal($q$)
        if $q > m$ then return true endif
        if Insert_Row($q$, $Q$)
          then return Extend_Transversal($q+1$)
          else
            $U \leftarrow Q + \{q\}$
            [$U$ is a dependent set of rows]
            $s \leftarrow q$
            if any row in $U$ marked then return false endif
              [This return means a row of B is in a dependent set.]
            for $r \in U$ do
              Reorganize $T$ to include row $s$ and exclude row $r$
              mark row $r$
              if Insert_Row($r$, $Q$)
              [Insert_Row is here being used to insert a row of B]
                then if Extend_Transversal($q+1$)
                  then return true
                    [We have found transversal]
                  endif
                endif
              [We were unable to extend transversal through a row of B, or
              failed to extend transversal through rest of matrix. Time to back-
              track.]
              unmark row $r$
              $s \leftarrow r$
              enddo
            return false
          endif
        endfunction Extend_Transversal
```

This algorithm handles the determination of the truth of (1.5). It can be extended to handle (1.3) without much difficulty, although the execution time can increase considerably. The extension consists of modifying the code that switches rows of **A** and **B** to switch rows of **A** and **G** instead, where the rows of **G** are selected based on the structure of the rows of **H** as follows. When Algorithm II selects a row of **B** for a switch, it should examine the structure of the corresponding row of **H**. For each nonzero in that row, the corresponding row of **G** should be tried in the transversal computation. (This requires an additional loop at each level of recursion.) A transversal will include entries from rows of **A** and **G**. Therefore, no more than one copy of a row of **G** should appear in a transversal. This restriction can be implemented by providing another global Boolean array to mark the rows of **G**, testing the mark before a switch, marking one when it is switched, and unmarking it when it is switched back.

3. Proof of method.

In this section we show that the algorithm works mathematically (this is not the same as a proof of the correctness of the program). We also present an interesting lemma and two corollaries. Before getting to the main result, let us examine the notion of structural rank in more detail.

If we know the structure of **A**, its numerical rank is a function of all of its nonzero entries. However, except for a set of values of measure zero, the numerical rank is invariant. The structural rank is this value. However, if the entries of **A** are functions of other variables, the numerical rank may be less. For example, the structural rank of a full 2 by 2 matrix is two. However, if $a_{11}=a_{12}a_{21}/a_{22}$, then the numerical rank cannot exceed one. We will call this the *constrained structural rank.* In this problem we are dealing with the matrix in (1.3) or (1.5) whose entries are functions of the values of the nonzeros in **A** and **B**. We will need the following two obvious results.

PROPOSITION 1. *An upper bound on the constrained structural rank of a matrix is the length of a maximum transversal.*

PROPOSITION 2. *A lower bound on the constrained structural rank of a matrix is the free length of any transversal, where the free length is the number of entries in a transversal whose values can be chosen independently of the values of any other entries of the matrix not in that transversal.*

For the previous example of a 2 by 2 matrix, the free length of either transversal is one.

THEOREM. *Algorithm* II *computes a value of true for* Extend_Transversal (1) *if and only if the index does not exceed two. In this case, the index is one if and only if no rows have been marked.*

Proof. We deal first with the index one case in which **A** is nonsingular. If the problem has index one, a full transversal exists and will be found by Algorithm I. Algorithm II duplicates the steps of Algorithm I in this case and will never mark a row. On the other hand, if Extend_Transversal returns true with no rows marked, it must have returned with a true value from Insert_Row(q, Q) every time, and never have entered the second phase of Extend_Transversal which marks rows. Hence, the algorithm has never found a linearly dependent set because Insert_Row was successful every time. Therefore, **A** is nonsingular and the index is one.

We now consider the case that the algorithm finds a full transversal but has marked some rows. Consequently, the index is two or greater, and **A** has linearly dependent rows that are identified by the algorithm each time it computes a false value for Insert_Row. We wish to show that the index is two. If **A** is rank deficient by r, then r different sets, **U**, of linearly dependent rows of **A** are identified by Extend_Transversal. Each set of dependent rows of **A** corresponds to a feasible nonzero structure of a row of **N**. The entries of **A** will make these entries of **N** nonzero almost everywhere. Therefore, the rows of **NB** can contain nonzeros in positions corresponding to those rows of **B** that are selected by the nonzeros in **N**.

However, it is not sufficient simply to examine the nonzero structure of **NB** to determine if (1.5) is satisfied, since its values cannot necessarily be chosen independently. We illustrate this point with the following example. Let **A** have the structure

$$\begin{pmatrix} 0 & 0 & x & 0 & 0 \\ x & x & x & 0 & 0 \\ x & x & x & 0 & 0 \\ x & 0 & 0 & 0 & 0 \\ x & 0 & 0 & 0 & 0 \end{pmatrix}$$

so that a possible structure for **N** might be

$$\begin{pmatrix} x & x & x & x & 0 \\ x & x & x & 0 & x \end{pmatrix},$$

but these entries in **N** are not independent since the submatrix comprising the second and third columns must necessarily be singular for **NA** to be zero.

We will show, however, that the values in the positions corresponding to the nonzeros in the rows of **B** selected by our algorithm are independently determined by entries of **B**. Note that the algorithm replaces a row of **A** with the corresponding row of **B** when it finds a dependent set. This is done before it looks for the next dependent set. Therefore, all subsequent dependent sets identified will not contain that row of **A**. Hence, the structure of **N** is such that each row has at least one nonzero entry in a column which contains zeros in all other rows (that is, each row has a column singleton). We identify one such element in each row and call it a *substitution entry.* If the algorithm terminates with the value true, it will have identified r dependent sets, replaced a row of **A** corresponding to each of those sets with a row of **B** in the position corresponding to the substitution entry, and found a transversal of length m for the combined matrix. The algorithm selects a row of **B** corresponding to each row of **NB**. The entries in a row of **NB** in the positions of the nonzero entries in the corresponding row of **B** can be independently chosen by varying the entries in the row of **B** without changing any other entries in **NB** because the substitution entry in **N** is the only nonzero in its column. The entries of **A** can all be freely chosen. Hence, the free length of the transversal is m, implying that (1.5) is satisfied. Thus we conclude that the structural index is two.

Finally we must show that if the index is two, the algorithm returns true, with some rows marked. If the rank of **N** is r, we can discard r dependent rows of **A**. The remaining rows of **A** plus the rows of **NB** must contain a transversal of length m for (1.5) to hold. Furthermore, there must be a transversal such that the submatrix formed by selecting the columns of **NB** corresponding to the columns of the transversal entries in **NB** is numerically nonsingular. (See Corollary 1 of the following lemma.) Corresponding to this transversal, we can make a $(1-1)$ correspondence between the r rows of **NB** and a set of r different rows of **B** such that the transversal entries in the rows of **NB** depend on an entry from the corresponding row of **B**. **N** must have an appropriate nonzero to pick up these entries of **B**. We will show that the algorithm will form this arrangement of rows of **A** and **B** unless it finds another arrangement that also gives a transversal of length m.

Consider the structure of an **N** that satisfies (1.5). We first arrange its rows so the column index of the last nonzero entry in each row is monotonically increasing. If this function is not strictly monotonically increasing, we replace a row with a linear combination of the adjacent rows to zero the last nonzero, and repeat this until the function is strictly monotonically increasing. This is equivalent to premultiplying **N** by a nonsingular matrix and does not affect (1.5). The first row of **N** now represents the first linear dependency of rows of **A**. It will be detected by the algorithm. The first row of **NB** must have a transversal entry since (1.5) is true. Suppose it is in column c. There must be a nonzero in column c of a row of **B** that corresponds to a nonzero in row 1 of **N**. Suppose the nonzero in **N** is $\mathbf{N}_{1r}$ (and the nonzero in **B** is $\mathbf{B}_{rc}$). Zero the entries in column r of all later rows in **N** by linear row operations. At some point, the algorithm will switch rows $\mathbf{A}_r$ and $\mathbf{B}_r$, and Insert_Row will be successful in finding a transversal through this row of **B**. This argument can be repeated for each linearly dependent set (row of **N**). The algorithm may choose a different column from c, but this path either will lead to another transversal or will fail, backtrack, and eventually choose c since the search is exhaustive. Therefore, if the index is two, the algorithm will find either the transversal of length m described above, or another. QED

It remains to show the following.

LEMMA. *If a matrix,* **A**, *is nonsingular but has a numerically singular r by r submatrix* **W**, *there must be a transversal of* **A** *that has no more than* $r-1$ *entries in* **W**.

Proof. Suppose **W** is the top left-hand corner of **A**. Expand det (**A**) in r by r minors using the first r columns of **A**. We have

$$\det(\mathbf{A}) = \sum_{x\in\mathbf{X}} (-1)^l \det(A_x)\det(A_y),$$

where $\{\mathbf{A}_x, x\in\mathbf{X}\}$ is the set of all r by r minors from the first r columns of **A**, $\mathbf{A}_y$ are the corresponding $m-r$ by $m-r$ minors from the last $m-r$ columns of **A**, and l is odd or even depending on which rows are in the minors. One of the summands consists of the top left-hand r by r minor, which is singular. Therefore, one of the other summands must be nonzero. Hence, the $\mathbf{A}_x$ term must have a transversal of length r, and the $\mathbf{A}_y$ term must have a transversal of length $m-r$, which together form a full transversal of **A**. No more than $r-1$ entries of the transversal of this $\mathbf{A}_x$ and none of the entries of the corresponding $\mathbf{A}_y$ are in **W**. QED

COROLLARY 1. *If* **W** *is a set of r rows of a nonsingular matrix* **A**, *a full transversal of* **A** *exists such that the submatrix of* **W** *containing the entries of that transversal is nonsingular.*

Proof. Suppose the contrary. Consider the set of all full transversals of **A**. Each of these have r entries in **W**. Let **W**′ be the smallest submatrix of **W** that contains all of the entries of all of these transversals in **W**. Every r by r minor of **W**′ is singular, so **W**′ is singular. The above lemma says that there exists another full transversal containing an entry not in **W**′, contrary to the definition of **W**′. QED

Another interesting corollary of this lemma follows.

COROLLARY 2. *If a matrix* **A** *is nonsingular but a submatrix* **W** *has numerical rank no greater than r, then there must exist a full transversal of* **A** *with no more than r entries in* **W**.

Proof. Without loss of generality, assume that **W** is in the top left-hand corner of **A** and has dimension p by q, where $p \geqq q \geqq r$. If $q = r$, then there is nothing to prove, since r columns can have no more than r entries in a transversal. Otherwise, $q > r$. Since **A** is nonsingular, it has a full transversal. Suppose no such transversal has less than $r+1$ entries in **W**. Choose a transversal with a maximum number of entries, s, not in **W**, and order the matrix so that the first $r+1$ columns of **W** contain transversal entries. Now expand the determinant of **A** by the minors of the first $r+1$ columns and the last $m-r-1$ columns, as in the lemma above. Since **A** is nonsingular, the determinants of at least one pair of corresponding minors are nonzero. The minors of the first $r+1$ columns either are drawn entirely from **W**, in which case their determinant is zero since **W** has rank r, or they include rows of **A** that are not part of **W**. One of the latter must be nonsingular, along with its corresponding minor from the last columns. Therefore, that pair has a full transversal of the $(r+1)$ by $(r+1)$ minor containing an entry of **A** not in **W** and a full transversal of the $(m-r-1)$ by $(m-r-1)$ minor. Together they form a full transversal of **A** with $s+1$ entries in **A** but not in **W**, contrary to the supposition. QED

REFERENCES

[1] C. W. GEAR AND L. R. PETZOLD, ODE *methods for the solution of differential/algebraic systems*, SIAM J. Numer. Anal., 21 (1984), pp. 716-728.

[2] K. BRENAN, *Stability and convergence of difference approximations for higher index differential/algebraic equations with applications to trajectory Control*, Ph.D. dissertation, Math. Dept., UCLA, Los Angeles, CA, 1983.

[3] S. L. CAMPBELL, *The numerical solution of higher index time varying singular systems of differential equations*, SIAM J. Sci. Stat. Comp., 6 (1985), pp. 334-348.

[4] P. LOTSTEDT AND L. R. PETZOLD, *The numerical solution of nonlinear differential equations arising with algebraic constraints*, Sandia Report SAND83-8877, Livermore, CA, Nov. 1983.

[5] I. S. DUFF, *On algorithms for obtaining a maximum transversal*, ACM Trans. Math. Soft., 7 (1981), pp. 315–330.

[6] R. F. SINCOVEC, A. M. ERISMAN, E. L. YIP AND M. A. EPTON, *Analysis of descriptor systems using numerical algorithms*, IEEE Trans. Automat. Control, Vol. AC-26 (1981), pp. 139–147.

[7] T. F. COLEMAN AND A. POTHEN, *The sparse null space basis problem*, Report TR 84-598, Dept. Comput. Sci., Cornell Univ., Ithaca, NY, July 1984.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 4, October 1986

011

SEQUENCE ALIGNMENTS WITH MATCHED SECTIONS*

JERROLD R. GRIGGS†, PHILIP J. HANLON‡ AND MICHAEL S. WATERMAN¶

Abstract. In molecular biology, two finite sequences are compared by displaying one sequence written over another in an alignment. The number of alignments of two sequences is related to the Stanton-Cowan numbers. This paper gives asymptotics for the number of alignments of two sequences of length n with matching sections of size at least b.

Key words. sequence alignments, generating functions, Stanton-Cowan numbers

AMS(MOS) subject classifications. 05A15, 92A10

Mathematics has played an important role in modern molecular biology in the area of sequence comparison. When nucleic acid (DNA or RNA) or protein sequences are determined, the question of relationships between sequences arises. Frequently two (or more) sequences are compared by dynamic programming or other methods to produce one or more sequence alignments which display one sequence written over another. When one letter (nucleotide in DNA) is written above another, they are presumed to have a common evolutionary ancestor. When a gap appears above or below a letter, the evolutionary event of insertion or deletion is assumed to have taken place. A review of methods to perform this analysis appears in Waterman [7].

An example of two different alignments of two sequences appears in Fig. 1(a) and 1(b) (Fitch and Smith [2]). The upper sequence is chicken β-hemoglobin messenger RNA (mRNA), nucleotides 115–171, and the lower sequence is chicken α-hemoglobin mRNA, nucleotides 118–156. These mRNA sequences are transcribed into hemoglobin protein molecules and are well known to have arisen from a common ancestor. In fact so many hemoglobin sequences are known that the alignment is presumed known, and the paper of Fitch and Smith is a study of the ability of various alignment algorithms to produce correct results.

As is easy to imagine, many ad hoc methods have arisen to align sequences. The most naive simply look at the sequences and perform the alignment visually. In order

```
(a) UUUGCGUCCUUUGGAAC CUCUCCAGCCCCA CUG C CAUCCUUGGCAA CC C CAUGG  UC
    UUU C  CC        CACU UC  G    AUCUGUCACA C   GGC  UCCGCUCA   AAUC

(b) UUUGCGUCCUUUGGAACCUCUCCAGCCCCAGUGCCAUCCUUGGCAACCCCAUGGUC
    UUUCCCCACUUCG  AUCU     GUCACACGGCUCCGCU      CAAAUC

(c) 11111111111111111011111111111110111010111111111111011010111110011
    11101001100000001111011001000011111111110100011100111111110001111

(d) 111111111111111111111111111111111111111111111111111111111
    11111111111110011110000011111111111111110001110000000111
```

FIG. 1. (a) *and* (b) *are two alignments of nucleotides* 115–171 *of chicken β-hemoglobin* mRNA (*upper*) *and nucleotides* 118–156 *of chicken α-hemoglobin* mRNA (*lower*). (c) *and* (d) *are* 0–1 *representations of* (a) *and* (b), *respectively.*

* Received by the editors July 29, 1985, and in revised form February 14, 1986.

† Department of Mathematics, University of South Carolina, Columbia, South Carolina 29208. This research was supported in part by the National Science Foundation.

‡ Department of Mathematics, California Institute of Technology, Pasadena, California 91125. This research was supported in part by the National Science Foundation.

¶ Departments of Mathematics and Molecular Biology, University of Southern California, Los Angeles, California 90089. This research was supported in part by the System Development Foundation.

to estimate the complexity of this task it is of interest to count the number of alignments for two sequences of two given lengths. There are previous results on this problem. H. T. Laquer [4] solves a more general recursion equation and relates the number of sequence alignments to the Stanton-Cowan numbers.

Frequently biologists find an alignment more believable when the matches occur in larger blocks. We will represent alignments as rows of 0's and 1's where a 1 indicates presence of a letter or nucleotide and a 0 indicates a gap. Figure 1(c) and 1(d) convert the alignments of Fig. 1(a) and 1(b) into these 0–1 rows. In this paper we count the alignments where the matching 1's must occur in blocks of b or more. In Fig. 1(a) and 1(c), $b=1$ while in Fig. 1(b) and 1(d), $b\leqq 3$.

Let $g(b, n)$ denote the number of alignments of two sequences of size n in which matching sections have size at least b. Equivalently, $g(b, n)$ is the number of $(0, 1)$-matrices with 2 rows and an unspecified number of columns such that both rows contain precisely n 1's, each column contains at least one 1, and columns with two 1's occur in adjacent sections of size b or more. We are interested in the asymptotic behavior of $g(b, n)$ for fixed b as $n\to\infty$, as a function of b.

Observe that alignments where no column sum equals 2 are simply permutations of n columns with a single 1 in row 1 and n columns with a single 1 in row 2. Those are satisfactory for any b. Thus for all b and n,

$$g(b, n)\geqq\binom{2n}{n}. \tag{1}$$

Applying Stirling's formula as $n\to\infty$ with b fixed,

$$g(b, n)\geqq((\pi n)^{-1/2})(4^n+o(1)) \quad \text{as } n\to\infty. \tag{2}$$

Further, note that $g(1, n)$ counts the total number of 2-sequence alignments. A generating function approach is successful for the general problem of $b\geqq 1$.

THEOREM 1. *Let $b\geqq 1$. Define*

$$h(x)=(1-x)^2-4x(x^b-x+1)^2$$

and let ρ be the smallest positive real root of $h(x)=0$. Then

$$g(b, n)\sim(\gamma_b n^{-1/2})D_b^n \quad \textit{as } n\to\infty,$$

where $D_b=\rho^{-1}$ and

$$\gamma_b=(\rho^b-\rho+1)(-\pi\rho h'(\rho))^{-1/2}.$$

Proof. Assume that b is fixed, $b\geqq 1$. Let $G(x)=\sum_{n\geqq 0} g(b, n)x^n$ denote the ordinary generating function for the numbers $g(b, n)$. In order to obtain $G(x)$ we first form the generating function $\phi_m(x)$ for the numbers of 2-sequence alignments in which there are precisely m columns each of the forms $\genfrac{}{}{0pt}{}{1}{0}$ and $\genfrac{}{}{0pt}{}{0}{1}$ and in which the columns $\genfrac{}{}{0pt}{}{1}{1}$ come in sections of at least b. As noted above, there are $\binom{2m}{m}$ ways to order the $2m$ columns with sum 1. This contributes a factor of $\binom{2m}{m}x^m$ to $\phi_m(x)$ since each row gets m 1's from these $2m$ columns. Next observe that there are $2m+1$ slots into which may be inserted either no $\genfrac{}{}{0pt}{}{1}{1}$ columns or at least b $\genfrac{}{}{0pt}{}{1}{1}$ columns. These slots precede, go between, and follow the $2m$ columns with one 1. So each such slot contributes a factor, call it $y=y(x)$, to $\phi_m(x)$, where

$$\begin{aligned} y=y(x)&=1+x^b+x^{b+1}+\cdots\\ &=1+(x^b)/(1-x)\\ &\Rightarrow y=(x^b-x+1)/(1-x). \end{aligned}$$

Hence,

$$\phi_m(x)=\binom{2m}{m}x^m y^{2m+1}. \tag{3}$$

We obtain (3) since each alignment coded by $\phi_m(x)$ is determined completely by the permutation of its columns with sum 1 and by the number s of $\genfrac{}{}{0pt}{}{1}{1}$ columns inserted into each slot. Such an alignment of size n contributes a term x^n to the sum $\phi_m(x)$.

The set of all 2-sequence alignments with columns $\genfrac{}{}{0pt}{}{1}{1}$ in groups of size at least b is the union over $m \geqq 0$ of the alignments enumerated by the series $\phi_m(x)$. Hence we obtain:

$$\begin{aligned}G(x)&=\sum_{m\geqq 0}\phi_m(x)\\&=\sum_{m\geqq 0}\binom{2m}{m}x^m y^{2m+1}\\&=y\sum_{m\geqq 0}\binom{2m}{m}(xy^2)^m.\end{aligned}$$

Applying the Binomial theorem,

$$G(x)=y(1-4xy^2)^{-1/2}.$$

Plugging in for y, we obtain

$$G(x)=(x^b-x+1)(h(x))^{-1/2},$$

where

$$h(x)=(1-x)^2-4x(x^b-x+1)^2$$

or

$$h(x)=1-6x+9x^2-4x^3-8x^{b+1}+8x^{b+2}-4x^{2b+1}.$$

Observe that $h(0)=1$ and $h(\frac{1}{4})=(\frac{3}{4})^2-((\frac{1}{4})^b+\frac{3}{4})^2<0$, so that h has a real root in $(0,\frac{1}{4})$. Let ρ be the smallest such root of h. The radius of convergence of $G(x)$ is determined by the roots of $h(x)$, so the following lemma implies that $G(x)$ has radius of convergence ρ.

LEMMA. *The unique root of* $h(x)$ *with the smallest modulus is* ρ, *and* ρ *is a single root of* $h(x)$.

Proof of Lemma. Let $z\in C$, $|z|\leqq\rho$, be a root of $h(z)$. We first show that in fact $|z|=\rho$ must hold. We have that

$$h(z)=(1-z)^2 4z\left(\frac{1}{4z}-\frac{(z^b-z+1)^2}{(1-z)^2}\right)=0.$$

Since $0<z<\frac{1}{4}$, it follows that

$$\frac{1}{4z}=\left(1+\frac{z^b}{1-z}\right)^2,$$

so that

$$\begin{aligned} \frac{1}{4\rho} \leqq \frac{1}{4|z|} &= \left|1+\frac{z^b}{1-z}\right|^2 \\ &\leqq \left(1+\frac{|z|^b}{1-|z|}\right)^2 \\ &\leqq \left(1+\frac{\rho^b}{1-\rho}\right)^2. \end{aligned}$$

Next we observe that because ρ is a root of h,

$$\frac{1}{4\rho} = \left(1+\frac{\rho^b}{1-\rho}\right)^2,$$

which implies that the inequalities above are all equalities. It follows that $|z| = \rho$. (This could have been deduced instead from the well-known fact that a series $f(z) = \sum_{n=0}^{\infty} a_n z^n$ with real coefficients $a_n \geqq 0$ and with radius of convergence $\rho > 0$ has a singular point at $z = \rho$ ([5]; confer, e.g., [3]).)

We next observe that

$$\left|1+\frac{z^b}{1-z}\right| = 1+\frac{|z|^b}{1-|z|},$$

where $|z| = \rho \in (0, \frac{1}{4})$ forces

$$1+\frac{z^b}{1-z} = 1+\frac{|z|^b}{1-|z|},$$

so that $1/4z = 1+(z^b/(1-z))$ is real and positive. Hence z itself is real and positive, which implies that z must be ρ. Thus ρ is the unique root with the smallest modulus.

One can then calculate that

$$h'(\rho) = (1-\rho)(-1-\rho^{-1}-4b\rho^{(2b-1)/2}+4\rho^{1/2}).$$

It follows easily from $\rho \in (0, \frac{1}{4})$ that $h'(\rho) < 0$. Therefore ρ is only a single root of $h(z)$. This completes the proof of the lemma.

Returning to the theorem, we define functions $s(x)$, $A(x)$, $B(x)$ by:

$$\begin{aligned} &h(x) = (\rho - x)s(x), \\ &A(x) = (x^b - x + 1)(s(x))^{-1/2}, \\ &B(x) = (\rho - x)^{-1/2}. \end{aligned}$$

Then we have that

$$G(x) = A(x)B(x).$$

Here $A(x)$ has radius of covergence $> \rho$ since it follows from the lemma that $s(x)$ has not root z with $|z| \leqq \rho$. Also, $B(x)$ has radius of convergence ρ. Again by the binomial theorem,

$$B(x) = (\rho - x)^{-1/2} = \rho^{-1/2}\left(1-\frac{x}{\rho}\right)^{-1/2} = \rho^{-1/2} \sum_{n \geqq 0} \binom{2n}{n}\left(\frac{x}{4\rho}\right)^n,$$

so that

$$B(x) = \sum_{n \geqq 0} b_n x^n$$

where

$$b_n = \rho^{-1/2}\binom{2n}{n}(4\rho)^{-n}.$$

It remains to observe that $(b_{n-1}/b_n) \to \rho$ as $n \to \infty$ to apply a theorem of Bender [1, Thm. 2] to $G(x) = A(x)B(x)$ to deduce that

$$g(b, n) \sim A(\rho)b_n \quad \text{as } n \to \infty.$$

Of course, to calculate $A(\rho)$, we are taking $s(\rho) = \lim_{x\to\rho} (h(x)/(\rho - x)) = -h'(\rho)$. The theorem now follows immediately.

Table 1 lists some values of D_b and γ_b to 4 or more places. These were computed on a hand computer, using Newton's method to find the root ρ for each b.

$$g(b, n) \sim (\gamma_b n^{-1/2})D_b^n \quad \text{as } n \to \infty,$$

where $D_b = \rho^{-1}$ and $\gamma_b = (\rho^b - \rho + 1)(-\pi\rho h'(\rho))^{-1/2}$.

For comparison, recall that from (2), for all b, $g(b, n) \geqq \binom{2n}{n} \sim (.5641896)n^{-1/2}4^n$ as $n \to \infty$. Table 1 also suggests what happens to D_b and γ_b as $b \to \infty$, which is straightforward to derive from the observation that as $b \to \infty$ the smallest root of $h(x)$, ρ, increases and approaches $\frac{1}{4}$:

TABLE 1

b	D_b	γ_b
1	5.8284	.57268
2	4.5189	.53206
3	4.1489	.54290
4	4.0400	.55520
5	4.0103	.56109
10	4.00001	.564183

COROLLARY. *As $b \to \infty$, $D_b \to 4$ and $\gamma_b \to \pi^{-1/2}$.*

Acknowledgment. We thank David Richman for his helpful observations.

REFERENCES

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

[2] W. FITCH AND T. SMITH, *Optimal sequence alignments*, Proc. National Academy of Science, 80 (1983), pp. 1382-1386.

[3] K. KNAPP, *Problem book in the theory of functions* I, Dover, New York, 1948, problem 11-3, p. 30.

[4] H. T. LAQUER, *Asymptotic limits for a two-dimensional recursion*, Stud. Appl. Math., 64 (1981), pp. 271-277.

[5] R. P. STANLEY, personal communication.

[6] R. G. STANTON AND D. D. COWAN, *Note on a 'square functional' equation*, SIAM Rev., 12 (1970), pp. 277-279.

[7] M. S. WATERMAN, *General methods of sequence comparison*, Bull. Math. Biol., 46 (1984), pp. 473-500.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 4, October 1986

012

CASCADE ADDITION AND SUBTRACTION OF MATRICES*

W. N. ANDERSON, JR.†, T. D. MORLEY‡ AND G. E. TRAPP§

Abstract. The cascade connection of electrical n-port networks motivates the cascade sum of matrices. The electrical network situation pertains to Hermitian positive semidefinite matrices, while in this work the cascade addition operation is also considered for arbitrary matrices. Various properties of the cascade sum are presented, including conditions which guarantee the existence of the cascade sum and the associativity of the cascade sum. A related operation, the cascade difference, is also treated. The underlying structure of the cascade operations is developed using the theory of the shorted operator.

Key words. cascade, Shur complement

AMS(MOS) subject classifications. 15A24, 15A45, 93A20

1. Introduction.

In electrical network theory, it is common to consider the concept of an *n-port network*, [11], [22]. Such a network will have n pairs of terminals; each pair is called a *port*. At each port a current and a voltage is defined. The collection of port currents will form a port current vector a; the port voltages form the port voltage vector α. The network then defines a linear operator $A: C^m \to C^m$, and the vector form of Ohm's law is $Aa = \alpha$. If all port currents are possible, that is, if the domain of the function A is C^m, then A is called the *impedance matrix* of the network. Here C^m is complex m-space.

When two n-port networks are interconnected, a new network is obtained. It is reasonable to ask if the new network will have an impedance matrix, and if the new impedance matrix will be a function of the impedance matrices of the interconnected networks. In many circumstances the answer to these questions is yes [5], [7], [18], [21]. In this paper we continue the study of these questions in the specific case of the cascade connection.

We are motivated by the *cascade* connection of networks, as illustrated in Fig. 1. The ports of each network are divided into sets, symbolized by ① and ②. Network A has n_1 ports in group ① and n_2 ports in group ②; network B has n_2 ports in group ① and n_3 ports in group ②. Each port in group ② of network A corresponds to a port in group ① of network B. The connection is such that the same voltage is measured

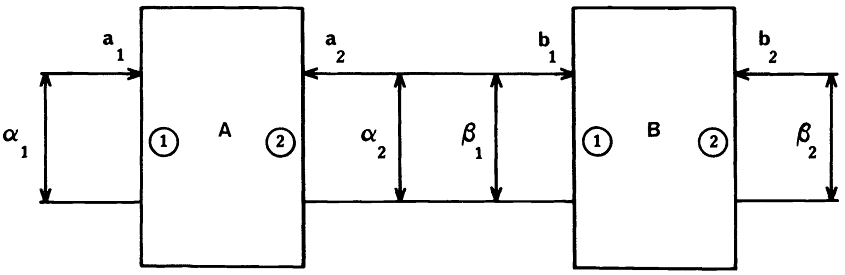


FIG. 1

* Received by the editors June 24, 1985, and in revised form January 15, 1986.

† Department of Mathematics and Computer Science, Fairleigh Dickinson University, Teaneck, New Jersey 07666.

‡ School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

§ Department of Statistics and Computer Science, West Virginia University, Morgantown, West Virginia 26506.

at corresponding ports, and the current into a port (in group 2) of network A is equal to the current out of the corresponding port of network B. If $n_1 = n_2 = n_3 = 1$, then the lines in Fig. 1 represent actual wires joining the networks. This cascade connection is important in both network analysis and synthesis [10], [16], [23]. It also arises in many apparently unrelated areas of science, see, for example, [19].

In terms of partitioned matrices, the network connection ensures the current equation $a_2 + b_1 = 0$ and the voltage equation $\alpha_2 = \beta_1$. These equations, together with Ohm's law, form the starting point of our investigations. The specific network model remains in the background for motivational purposes.

In our formulation of the network equations we implicitly assume that the connection is such that the currents at each of interconnected networks are defined as port currents. That is, the current into one terminal of a port must be equal to the current out of the other terminal of the same port. This is a standard assumption in the theory of n-port networks; physically this amounts to assuming that the connection includes the use of ideal transformers at each port. Without this condition the impedance matrix of the connection would be a function of the networks themselves, rather than being a function of the impedance matrices of the connected matrices [16]. Thus our mathematical theory applies only to the restricted physical setting of networks interconnected with ideal transformers.

In this paper we first consider general complex matrices, without discussing any particular physical realizations. We derive conditions for the existence of a cascade sum and differences, and develop some other properties of the sum. However, in any specific physical instance of the cascade connection, the class of matrices will be restricted. We will consider the special cases of positive semidefinite Hermitian matrices (corresponding to resistive networks), and almost right definite matrices (corresponding to passive networks). Other areas of interest include positive real matrices and the applications of cascade subtraction to classical network synthesis problems, see [17]. One could also consider the special cases arising in other physical settings, including operators acting on continuous (hence infinite dimensional) systems or connections of infinite numbers of operators. Examples of infinite networks may be found in the following references: [4], [10], [24].

We now present the outline of this paper. In § 2, we define our environment and present the appropriate preliminary information. Section 3 contains detailed information about the shorted operator (Schur Complement), which is fundamental to our presentation of cascade addition. In this section, we review and amplify some known results, and we present some new results. Section 4 deals with the general questions of the existence of the cascade sum of matrices, the existence of the cascade difference of matrices and cascade associativity. In § 5 we restrict the cascade operation to the case of positive semidefinite matrices. In § 6, we consider two topics motivated by electrical network theory, duality and the chain matrix, and show how they apply to the cascade sum. Section 7 deals with passive networks. In it we briefly discuss the cascade sum of almost right definite matrices. Finally, in § 8 we conclude by summarizing this work and briefly mentioning areas of future work.

2. Preliminaries. We consider vectors and linear operators defined on finite-dimensional complex inner product spaces. For vectors x and y, we use $\langle x, y\rangle$ to denote the inner product. For a linear operator A, we use the adjoint A defined by $\langle Ax, y\rangle = \langle x, A^*y\rangle$ for all vectors x and y. If $A = A^*$ we say that A is *Hermitian*.

When appropriately motivated by the network model, we will speak of *current* vectors, for which we will use lower case Latin letters, and *voltage* vectors, for which we will use lower case Greek letters. The terms "current" and "voltage" are not given any mathematical meaning; the vectors are merely assumed to lie in spaces appropriate to the context in which they appear.

Two special cases of matrices will be important. If A is a Hermitian matrix such that $\langle Aa, a\rangle \geqq 0$ for all vectors a, then A is said to be *positive semidefinite.* If A and B are positive semidefinite matrices, then we write $A \geqq B$ when $A-B$ is positive semidefinite. It is well known that for a positive semidefinite matrix the quadratic form $\langle Aa, a\rangle = 0$ only if $Aa = 0$.

For a matrix A, we let Range (A) and Kernel (A) denote the appropriate subspaces, and rank (A) and nullity (A) their dimensions.

For a matrix A, the symbol A^- denotes any matrix such that $AA^-A = A$; such a matrix is called a 1-inverse. A 1-inverse exists for any matrix A, but it will not be unique unless A is actually invertible. However, some important matrix expressions are independent of the 1-inverses appearing therein. A number of elementary results concerning 1-inverses are summarized below; where proofs are omitted they may be found in standard textbooks, for example, see [20].

LEMMA 1. *Let A be a matrix and A^- a 1-inverse. Then*

(a) *The product AA^- is a projection, and* Range $(AA^-) =$ Range (A).

(b) *The system $AX = B$ has a solution if and only if* Range $(B) \subset$ Range (A). *In this case $X = A^-B$ is a solution.*

(c) *Let B and C be nonzero matrices. Then the product BA^-C is independent of the choice of A^- if and only if* Range $(C) \subset$ Range (A), *and* Range $(B^*) \subset$ Range (A^*).

For a matrix A one choice of a 1-inverse is the Moore-Penrose pseudo-inverse A^+. The following well-known properties are important for this paper; see [20] for proofs.

LEMMA 2. *Let A be a matrix and A^+ its Moore-Penrose pseudo-inverse: then*

(a) *A^+ is the unique 1-inverse that satisfies $A^+AA^+ = A^+$; A^+A is Hermitian; AA^+ is Hermitian.*

(b) *AA^+ is the Hermitian projection onto* Range (A).

(c) *A^+A is the Hermitian projection onto* Range (A^*).

(d) *If A is positive semidefinite then A^+ is positive semidefinite also, and* Range $(A^+) =$ Range (A).

(e) $(A^*)^+ = (A^+)^*$.

(f) Range $(A^*) =$ Range (A^+).

(g) Kernel $(A^*) =$ Kernel (A^+).

LEMMA 3. *Let A be a positive semidefinite matrix, and Q a matrix with* Range $(Q) =$ Range (A). *Let B be a positive semidefinite matrix with* Range $(B) \cap$ Range $(A) = 0$. *Then*

(a) Range $(Q^*A^+Q) =$ Range (Q^*).

(b) $(Q^*A^+Q)^+ = Q^+AQ^{*+}$.

(c) $Q^*A^+Q = Q^*(A+B)^+Q$.

Proof. For part a, if x is a vector such that $Q^*A^+Qx = 0$, then $0 = \langle Q^*A^+Qx, x\rangle = \langle A^+Qx, Qx\rangle$. Thus $A^+Qx = 0$ since A is positive semidefinite. But then $Qx = 0$ since Range $(Q) =$ Range (A^+), and A^+ is injective on its range.

Conversely, if $Qx = 0$ then $Q^*A^+Qx = 0$. Therefore Kernel $(Q^*A^+Q) =$ Kernel (Q); since both operators are Hermitian, their ranges are equal also.

For part (b), it is straightforward to verify that the conditions of Lemma 2(a) hold.

The proof of part (c) will be deferred until after the proof of Lemma 12. QED

Throughout the paper we will be using partitioned matrices; in all cases we assume without comment that the dimensions of the blocks are consistent with the indicated partition. The partitioning of the ports in Fig. 1 gives rise to a natural partitioning of the impedance matrices, so that

$$(1)\qquad A=\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad \text{and} \quad B=\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

A similar partition is used for the cascade sum C of A and B, and in connection with the discussion of the shorted operator. The voltage and current vectors are also partitioned to be compatible with the matrix partitions. In order for the equations defining the cascade sum to be meaningful, it is necessary for A_{22} and B_{11} to have the same dimension; the dimensions of A and B are otherwise arbitrary. Thus we can say that A is an (m_1+m_2) by (n_1+n_2) matrix, and B is an (m_2+m_3) by (n_2+n_3) matrix; the cascade sum C is then (m_1+m_3) by (n_1+n_3). In particular, it is not necessary for A and B to be square, although only square matrices are meaningful in the network model.

In our treatments of the special case of positive semi-definite matrices we will always assume that $m_i=n_i$, $i=1, 2, 3$.

3. The shorted operator. The *shorted operator*, also called the *Schur Complement*, is fundamental to the theory of matrix operations induced by network connections, see [5] and [7]. In this section we develop those portions of the theory necessary for studying the cascade connection; more extensive treatments are contained in the references: [4], [9], [12], [22]. Theorem 4 and its corollaries are well known; the treatment here is different because of the emphasis on the currents and voltages. Theorems 7 and 13 appear to be new; their corollaries yield new arguments for old results and serve as an introduction to the arguments to be used in solving cascade subtraction problems. We will first treat the case of general matrices, and then turn to the special case of positive semidefinite matrices.

For the electrical network background of the shorted operator, consider the n-port network A in Fig. 2. The ports of A are partitioned into two sets. The ports in the second set will be shorted, and an input current will be applied to the ports in the first set. Thus the impedance matrix is naturally partitioned as in (1). The current vector a and the voltage vector α will be partitioned in a similar manner.

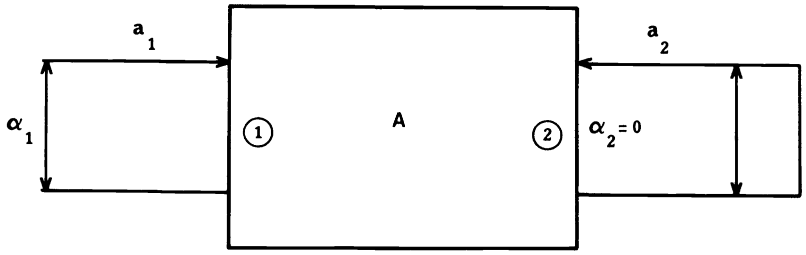


FIG. 2

Thus for a current a_1 we seek a current a_2 such that the voltage is zero (because of the short). The map from the input current a_1 to the voltage α_1 will define the impedance matrix of the shorted network; this map is the shorted operator of A.

The algebraic definition of the Schur Complement is directly analogous to the network model, but is somewhat more general because the matrices involved are not assumed to be square. For convenience, we will use a slightly different notation for the blocks of the partitioned matrix.

Given a partitioned matrix

$$M=\begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

and a vector a, we wish to find a vector b such that

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix}=\begin{bmatrix} \alpha \\ 0 \end{bmatrix}. \tag{2}$$

If

(i) for each vector a there exists a vector b such that (1) holds;

(ii) the vector α is uniquely determined;

then we say that the matrix M is *complementable*, and define $\mathcal{S}(M)a=\alpha$.

We note that the operator $\mathcal{S}(M)$ is linear, since in (2) a linear combination of a's and the same linear combination of b's will result in the same linear combination of α's. Thus we may use a matrix for $\mathcal{S}(M)$.

THEOREM 4. *The matrix M is complementable if and only if* Range $(C) \subset$ Range (D) *and* Range $(B^*) \subset$ Range (D^*).

Proof. Given a, we must solve $Ca+Db=0$ for b. But Ca may be any vector in Range (C), and thus we must have Range $(C) \subset$ Range (D). Conversely, if the condition holds we may always solve for b.

Now suppose that $a=0$. We must ensure that $\alpha=0$. But

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} 0 \\ b \end{bmatrix}=\begin{bmatrix} \alpha \\ 0 \end{bmatrix}=\begin{bmatrix} Bb \\ Db \end{bmatrix}.$$

Since $Db=0$, the necessary and sufficient condition is that Kernel $(D) \subset$ Kernel (B), equivalently Range $(B^*) \subset$ Range (D^*). In view of Theorem 4, our definition of complementable is equivalent to that given in [9] and [15].

COROLLARY 5. *If M is complementable then M^* is complementable and $\mathcal{S}(M^*)=(\mathcal{S}(M))^*$.*

Proof. The range conditions for M^* are the same as those for M; therefore M^* is complementable.

Now let a and a' be vectors. We need to show that $\langle \mathcal{S}(M)a, a'\rangle=\langle a, \mathcal{S}(M^*)a'\rangle$. But if

$$M\begin{bmatrix} a \\ b \end{bmatrix}=\begin{bmatrix} \alpha \\ 0 \end{bmatrix} \quad \text{and} \quad M^*\begin{bmatrix} a' \\ b' \end{bmatrix}=\begin{bmatrix} \alpha' \\ 0 \end{bmatrix}$$

we have

$$\begin{aligned}
\langle \mathcal{S}(M)a, a'\rangle=\langle \alpha, a'\rangle&=\left\langle \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} a' \\ b' \end{bmatrix}\right\rangle \\
&=\left\langle \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} A^* & C^* \\ B^* & D^* \end{bmatrix}\begin{bmatrix} a' \\ b' \end{bmatrix}\right\rangle \\
&=\left\langle \begin{bmatrix} a \\ b' \end{bmatrix}, \begin{bmatrix} \alpha' \\ 0 \end{bmatrix}\right\rangle \\
&=\langle a, \mathcal{S}(M^*)a'\rangle. \qquad \text{Q.E.D.}
\end{aligned}$$

COROLLARY 6. *If the matrix M is complementable then for any choice of the* 1*-inverse* D^-

$$\mathcal{S}(M)=A-BD^-C. \tag{3}$$

Proof. Since Range $(C)\subset$ Range (D), we can let $b=-D^-Ca$ for any choice of D^-. Then $Db=DD^-Ca=-Ca$. Then $\alpha=(A-BD^-C)a$ is independent of the choice of D^- since Range $(B^*)\subset$ Range (D^*). Q.E.D.

The next theorem about partitioned matrices is directly related to our work in subsequent sections, and it also has an interesting relationship to parallel addition.

THEOREM 7. *The equation*

$$\mathcal{S}\begin{bmatrix} A & B \\ C & X \end{bmatrix}=0$$

has a solution X if and only if

(i) Range $(A)\subset$ Range (B) *and* Range $(A^*)\subset$ Range (C^*),

(ii) nullity $(B)\geqq$ rank $(C)-$ rank (A),

(iii) nullity $(C^*)\geqq$ rank $(B^*)-$ rank (A^*).

Proof. First suppose that there is a solution; that there is a D such that Range $(C)\subset$ Range (D), Range $(B^*)\subset$ Range (D^*), and $A=BD^-C$. Then (i) is obviously satisfied. Since Range $(C)\subset$ Range (D) implies rank $(C)=$ rank (D^-C), (ii) follows from rank $(A)=$ rank $(D^-C)-$ dim (Range $(D^-C)\cap$ Kernel $(B))\geqq$ rank $(C)-$ nullity (B). Inequality (iii) follows similarly.

For sufficiency, we note that Range $(A)\subset$ Range (B) and Range $(A^*)\subset$ Range (C^*) imply that the matrix $Y=B^+AC^+$ satisfies $A=BYC$. It is clear that rank $(A)=$ rank $(B^+A)=$ rank (YC). From rank $(C)=$ rank $(YC)+$ dim (Kernel $(Y)\cap$ Range $(C))$ and (ii) it follows that dim (Kernel $(Y)\cap$ Range $(C))\leqq$ nullity (B). Thus there is a matrix L with Kernel $(L)=($Kernel $(Y)\cap$ Range $(C))^\perp$ and Range $(L)\subset$ Kernel (B). In a similar manner it follows from (iii) that there exists a matrix M with Kernel $(M^*)=($Kernel $(Y^*)\cap$ Range $(B^*))^\perp$, and Range $(M^*)\subset$ Kernel (C^*). We will show that $X=(Y+L+M)^+$ is a solution to the shorted subtraction problem.

It is clear that $BX^+C=A$, since $BL=0$ and $MC=0$ by construction. It remains to verify the range conditions of Theorem 4. We will show that Kernel $(X^*)\subset$ Kernel (C^*); the other condition will follow by a similar argument.

First suppose x is a vector such that $(Y+L)x=0$. Then $Yx=-Lx\in$ Range $(Y)\cap$ Range $(L)\subset$ Range $(B^*)\cap$ Kernel $(B)=0$. Thus Kernel $(Y+L)=$ Kernel $(Y)\cap$ Kernel (L). Similarly, if $(Y+L+M)x=0$, then $(Y+M)x=-Lx\in$ Range $(Y+M)\cap$ Range (L). But Range $(M)=$ Kernel $(Y^*)\cap$ Range $(B^*)\subset$ Range (B^*), and Range $(Y)\subset$ Range (B^*), so that Range $(Y+M)\cap$ Range $(L)=0$. Thus $(Y+M)x=-Lx=0$. Then $Yx=-Mx\in$ Range $(Y)\cap$ Range $(M)=$ Range $(Y)\cap$ Kernel $(Y^*)\cap$ Range $(B^*)=0$. Therefore Kernel $(Y+L+M)=$ Kernel $(Y+L)\cap$ Kernel $(M)\subset$ Kernel $(Y+L)$.

Next we will show that Kernel $(Y+L)=$ Kernel (C^*). Suppose $C^*x=0$. Then $Yx=B^+AC^+x=0$. Moreover, from Range $(C)\supset$ Range $(C)\cap$ Kernel (Y) we have Kernel $(C^*)\subset($Range $(C)\cap$ Kernel $(Y))^\perp=$ Kernel (L). Thus Kernel $(C^*)\subset$ Kernel $(Y)\cap$ Kernel $(L)\subset$ Kernel $(Y+L)$. Consider now an $x\in$ Kernel $(Y+L)-$ Kernel (C^*). That is $x\in$ Kernel $(Y+L)$ and $x\in$ Range (C), so that $x=Cz$ and $(Y+L)Cz=0$. From the previous paragraph, we have $Cz\in$ Kernel $(Y)\cap$ Kernel (L);

therefore $Cz \in \text{Kernel}(L) \cap (\text{Kernel}(Y) \cap \text{Range}(C))^{\perp} = 0$. Thus $x = 0$ and Kernel $(Y+L) = \text{Kernel}(C^*)$.

Thus Kernel (X^*) = Kernel $(Y+L+M)^{+*}$ = Kernel $(Y+L+M) \subset$ Kernel $(Y+L) = \text{Kernel}(C^*)$. Q.E.D.

For matrices A and B, the *parallel sum* of A and B, denoted by A: B, is defined to be the Schur Complement of the partitioned matrix

$$\begin{bmatrix} A & A \\ A & A+B \end{bmatrix}.$$

The parallel sum is another example of a matrix operation induced by a network connection, see [5]. For appropriate operators A and B an equivalent formula is $A\colon B = (A^{-1}+B^{-1})^{-1}$. Basic results concerning the parallel sum of positive semidefinite matrices are given in [4], and for general matrices in [21].

COROLLARY 8. *Let A and B be m by n matrices. The equation $A\colon X = C$ has a solution X if and only if*

(i) Range $(C) \subset$ Range (A) *and* Range $(C^*) \subset$ Range (A^*),

(ii) rank $(A-C) \geqq 2$ rank $(A) - \min(m, n)$.

Proof. The equation $A\colon X = C$ is equivalent to the equation

$$\mathcal{S}\begin{bmatrix} A-C & A \\ A & A+X \end{bmatrix} = 0.$$

By Theorem 7 there is a solution if and only if

(i) Range $(A-C) \subset$ Range (A) and Range $(A-C)^* \subset$ Range (A^*),

(ii) nullity $(A) \geqq$ rank $(A) -$ rank $(A-C)$,

(iii) nullity $(A^*) \geqq$ rank $(A^*) -$ rank $(A-C)^*$.

It is an easy exercise to see that the two sets of conditions are the same. This result was originally proved in [21]. Q.E.D.

When dealing with positive semidefinite matrices, we will assume that in partition (1) the matrix A and minor A_{22} are square. We will use the term *shorted operator* for this special case of the Schur Complement.

To begin our treatment of the positive semidefinite case, let us first recall four lemmas which are proved in [1] and [4].

LEMMA 9. *Let A be a positive semidefinite matrix, partitioned as in* (1), *with A_{22} square, and let S be the subspace corresponding to the first block in the partition. Let $\mathcal{M}(A, S)$ be the set of positive semidefinite matrices X such that $X \leqq A$ and* Range $(X) \subset S$. *Then A is shortable and*

(a) *$\mathcal{S}(A)$ is the maximum element of $\mathcal{M}(A, S)$.*

(b) Range $(\mathcal{S}(A)) =$ Range $(A) \cap S$.

(c) Range $(A - \mathcal{S}(A)) \cap S = 0$.

We note that in Lemma 9 the shorted operator is considered as an operator on the larger space, whereas the formula (3) gives an operator acting on S only. In the positive semidefinite case that we are considering here, $\mathcal{S}(A)$ will be zero on $S^{\perp}$, and thus no confusion should result from the identification of the two operators.

LEMMA 10. *Let A and B be positive semidefinite matrices, with $A \leqq B$. Then $\mathcal{S}(A) \leqq \mathcal{S}(B)$.*

LEMMA 11. *Let A be a Hermitian matrix, partitioned as in* (1). *Then A is positive semidefinite if and only if A_{22} and $\mathcal{S}(A)$ are positive semidefinite.*

LEMMA 12. *Let A and B be positive semidefinite matrices. Then $A\colon B$ exists and* Range $(A\colon B) =$ Range $(A) \cap$ Range (B).

We can now complete the proof of Lemma 3. Since Range $(A) \cap$ Range $(B) = 0$, we have $A: B = 0$ by Lemma 12. Then $Q^*A^+Q = Q^*A^+AA^+Q = Q^*A^+(A - A: B)A^+Q = Q^*A^+A(A+B)^+AA^+Q = Q^*(A+B)^+Q$. The last equality follows from Lemma 2 part (b) and the equality Range (A) = Range (Q). Q.E.D.

THEOREM 13. *Let A be a positive semidefinite matrix and B a matrix. Then there exists a matrix X such that*

$$M = \begin{bmatrix} A & B \\ B^* & X \end{bmatrix}$$

is positive semidefinite and $\mathscr{S}(M) = 0$ if and only if Range (A) = Range (B).

Proof. Since M is positive semidefinite, we must have Range $(B) \subset$ Range (A). In order to have $\mathscr{S}(M) = 0$, by Theorem 7 we must have Range $(A) \subset$ Range (B), and thus equality must hold.

Conversely, if the range equality holds, then we take $X = B^*A^+B$. Lemma 3 part (b) yields that $X^+ = B^+AB^{*+}$; and by a straightforward computation we see that $\mathscr{S}(M) = 0$. Q.E.D.

COROLLARY 14. *Let A and C be positive semidefinite matrices. Then there is a positive semidefinite matrix X such that A: $X = C$ if and only if $A - C$ is positive semidefinite and* Range $(A - C)$ = Range (A).

Proof. If the hypotheses hold, then Theorem 13 ensures the existence of a positive semidefinite Y such that

$$\mathscr{S}\begin{bmatrix} A-C & A \\ A & Y \end{bmatrix} = 0.$$

Choose $Y = A(A-C)^+A$. From $A - C \leqq A$, we have $(A-C)^+ \geqq A^+$, since the ranges on both sides are equal. Then $Y = A(A-C)^+A \geqq AA^+A = A$. Therefore we may write $Y = A + X$ with X positive semidefinite. Conversely, if A: $X = C$, then

$$\mathscr{S}\begin{bmatrix} A-C & A \\ A & A+X \end{bmatrix} = 0$$

and Theorem 13 applies. Q.E.D.

4. The cascade sum. The fundamental equations for the cascade sum are motivated by the physical model, as illustrated in Fig. 1. In this section we derive the conditions for the existence of the cascade sum and study some of its properties. Although we will continue to use the physical terminology of currents and voltages, all of the proofs will be purely algebraic. We will use the matrix partitioning conventions described in § 2.

Given matrices A and B, and a vector c, we seek vectors a and b such that

$$(4)\quad \left.\begin{array}{ll} \text{(i)} & a_1 = c_1, \\ \text{(ii)} & b_2 = c_2, \\ \text{(iii)} & a_2 + b_1 = 0, \end{array}\right\} \quad \text{(current conditions)},$$

$$\text{(iv)}\quad A_{21}a_1 + A_{22}a_2 = B_{11}b_1 + B_{12}b_2, \qquad \text{(voltage condition)}.$$

Given current vectors a and b satisfying (i)-(iv), we then define the voltage vectors α and β by $\alpha = Aa$, and $\beta = Bb$. The vector γ is then defined by $\gamma_1 = \alpha_1$, and $\gamma_2 = \beta_2$. We note that equation (iv) could also have been written $\alpha_2 = \beta_1$.

DEFINITION. If for all vectors c there exist vectors a and b satisfying (4), and if moreover the vector γ is uniquely determined by c, then we say that the matrices A and B are *cascade summable.* Since the relationship between c and γ is linear, there is a matrix C such that $Cc=\gamma$. The matrix C is called the *cascade sum* of A and B, and we write $C=A\circ B$.

In order to discuss existence of the cascade sum, let us introduce the matrix Q defined by the following matrix multiplication expression:

$$Q=\begin{bmatrix} I & 0 & 0 & 0\\ 0 & 0 & 0 & I\\ 0 & I & -I & 0\end{bmatrix}\begin{bmatrix} A_{11} & A_{12} & 0 & 0\\ A_{21} & A_{22} & 0 & 0\\ 0 & 0 & B_{11} & B_{12}\\ 0 & 0 & B_{21} & B_{22}\end{bmatrix}\begin{bmatrix} I & 0 & 0\\ 0 & 0 & I\\ 0 & 0 & -I\\ 0 & I & 0\end{bmatrix}.$$

Performing the matrix multiplications yields that Q may be rewritten as follows:

$$(5)\qquad Q=\begin{bmatrix} A_{11} & 0 & A_{12}\\ 0 & B_{22} & -B_{21}\\ A_{21} & -B_{12} & D\end{bmatrix}\quad \text{where } D=A_{22}+B_{11}.$$

The construction of this matrix Q is indicated by the general theory of matrix operations induced by network connections [5]. For the present purposes, we observe that equations (4) may be rewritten as

$$Q\begin{bmatrix} a_1\\ b_2\\ x\end{bmatrix}=\begin{bmatrix} \gamma_1\\ \gamma_2\\ 0\end{bmatrix}\quad \text{where } x=a_2=-b_1.$$

From the definition of complementability we see that the matrices A and B will be cascade summable if and only if Q is complementable to the subspace corresponding to the first two blocks in the partitioned form of the matrix in (5). Our fundamental existence theorem is then an instance of Theorem 4.

THEOREM 15. *The matrices A and B are cascade summable if and only if* Range $([A_{21}, -B_{12}])\subset$ Range (D) *and* Range $([A^*_{12}, -B^*_{21}])\subset$ Range (D^*).

Proof. Apply Theorem 4 to the matrix Q. Q.E.D.

THEOREM 16. *If the matrices A and B are cascade summable, then the cascade sum $A\circ B$ is given by*

$$(6)\qquad A\circ B=\begin{bmatrix} A_{11}-A_{12}D^{-}A_{21} & A_{12}D^{-}B_{12}\\ B_{21}D^{-}A_{21} & B_{22}-B_{21}D^{-}B_{12}\end{bmatrix}.$$

Proof. The formula is a direct application of (3) to the matrix Q. Q.E.D.

We note that the formula for $A\circ B$ is independent of the choice of the 1-inverse D^{-}. It is also clear that except in the trivial cases where $[A_{21}, -B_{12}]=0$ or $[A^*_{12}, -B^*_{21}]=0$, the matrices A and B will be cascade summable if and only if the right-hand side of (6) is independent of the choice of 1-inverse D.

In Fig. 3 we show the cascade connection of three networks. From the physical model one would expect the cascade sum to be associative, since it should make no difference whether the wires in the left connection or the right connection are soldered first. In the algebraic setting, it proves convenient to give a separate definition for the cascade sum of three matrices.

We say that the three matrices A, B and C are *tri-cascadable* if for every vector d there are vectors a, b and c such that

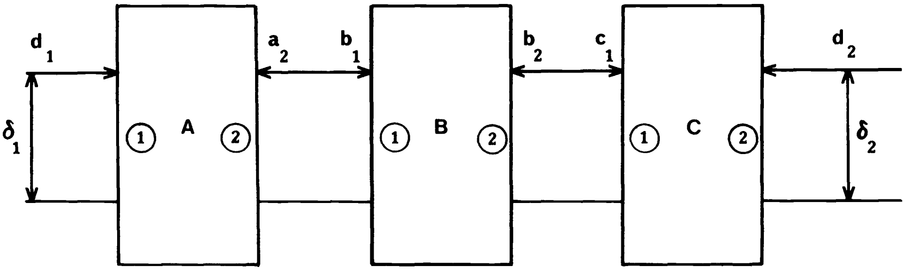


FIG. 3

$$\left.\begin{aligned} a_1 &= d_1, \\ c_2 &= d_2, \\ a_2 + b_1 &= 0, \\ b_2 + c_1 &= 0, \end{aligned}\right\} \quad \text{(current conditions),}$$

$$\left.\begin{aligned} A_{21}a_1 + A_{22}a_2 &= B_{11}b_1 + B_{12}b_2, \\ B_{21}b_1 + B_{22}b_2 &= C_{11}c_1 + C_{12}c_2, \end{aligned}\right\} \quad \text{(voltage conditions),}$$

and if the vectors $\delta_1 = A_{11}a_1 + A_{12}a_2$ and $\delta_2 = C_{21}c_1 + C_{22}c_2$ are uniquely determined by d.

If the matrices are tri-cascadable, we then have a matrix $\circ(A, B, C)$ such that $\delta = \circ(A, B, C)d$.

Proceeding analogously to the case of two matrices, we introduce vectors $x = a_2 = -b_1$, and $y = b_2 = -c_1$, and rewrite the tri-cascade equations

$$\begin{bmatrix} A_{11} & 0 & A_{12} & 0 \\ 0 & C_{22} & 0 & -C_{21} \\ A_{21} & 0 & D & -B_{12} \\ 0 & -C_{12} & -B_{21} & E \end{bmatrix} \begin{bmatrix} a \\ c \\ x \\ y \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ 0 \\ 0 \end{bmatrix}, \tag{7}$$

where $D = A_{22} + B_{11}$, and $E = B_{22} + C_{11}$. The conditions for tri-cascadability are the same as those for the matrix on the left-hand side of (7) to be complementable to the subspace corresponding to the first two blocks of the partition.

LEMMA 17. *Let A and B be cascade summable matrices. Then $(A \circ B) \circ C$ exists if and only if*·$\circ(A, B, C)$ *exists; if both exist they are equal.*

Proof. Assume that $(A \circ B) \circ C$ exists. Then for vectors d_1 and d_2 there exists a vector y such that $(A \circ B)_{21}d_1 + (A \circ B)_{22}y = -C_{11}y + C_{12}d_2$. Moreover the vectors $\delta_1 = (A \circ B)_{11}d_1 + (A \circ B)_{12}y$ and $\delta_2 = -C_{21}y + C_{22}d_2$ do not depend on y.

By the existence of $A \circ B$ (with d_1 and y replacing c_1 and c_2 in (4)) we know that there exists a vector x such that $A_{21}d_1 + A_{22}x = -B_{11}x + B_{12}y$. By the definition of $A \circ B$ we also have $(A \circ B)_{21}d_1 + (A \circ B)_{22}y = -B_{21}x + B_{22}y$, and therefore $-B_{21}x + B_{22}y = -C_{11}y + C_{12}d$. The x and y we have just constructed satisfy (7), with d_1, d_2 replacing a, c. Since δ_1 and δ_2 are uniquely determined, we have shown that $\circ(A, B, C)$ exists and is equal to $(A \circ B) \circ C$.

Conversely, assume that $\circ(A, B, C)$ exists. Then there exist x and y with $A_{21}d_1 + A_{22}x = -B_{11}x + B_{12}y$ and $-B_{21} + B_{22}y = -C_{11}y + C_{12}d_2$. Moreover $\delta_1 = A_{11}d_1 + A_{21}x$ and $\delta_2 = -C_{21}y + C_{22}d_2$ are uniquely determined. Again we use the definition of $A \circ B$, applied to the currents d_1 and y, to see that $(A \circ B)_{21}d_1 + (A \circ B)_{22}y = -C_{11}y + C_{12}d_2$ and that $\delta_1 = (A \circ B)_{11}d_1 + (A \circ B)_{12}y$. Since we still have the uniqueness of δ_1 and δ_2, we have proved that $(A \circ B) \circ C$ exists and equals $\circ(A, B, C)$. Q.E.D.

LEMMA 18. *Let B and C be cascade summable matrices. Then $A \circ (B \circ C)$ exists if and only if $\circ(A, B, C)$ exists. If both exist then they are equal.*

Proof. The proof is essentially the same as for the previous lemma. Q.E.D.

THEOREM 19. *Let A, B and C be matrices such that $(A \circ B) \circ C$ and $A \circ (B \circ C)$ both exist. Then $(A \circ B) \circ C = A \circ (B \circ C)$.*

Proof. By Lemmas 17 and 18, both expressions equal $\circ(A, B, C)$. Q.E.D.

Our next theorem deals with the cascade associativity question for just one matrix. When is $A \circ (A \circ A) = (A \circ A) \circ A$? We will note in the final section of this paper that the existence of the cascade limit operation is an interesting research question, and that the result below provides a foundation for further analysis of this question.

THEOREM 20. *Let A be a matrix. If $A \circ A$ and $A \circ (A \circ A)$ exist, then $(A \circ A) \circ A$ exists and $(A \circ A) \circ A = A \circ (A \circ A)$.*

Proof. Since $A \circ A$ and $A \circ (A \circ A)$ both exist, Lemma 18 implies that $\circ(A, A, A)$ exists. Using Lemma 17, we then see that $(A \circ A) \circ A$ exists and also equals $\circ(A, A, A)$. Q.E.D.

We close our discussion of associativity by remarking that one can construct 2 by 2 matrices A, B and C such that $A \circ (B \circ C)$ exists but $(A \circ B) \circ C$ does not. One can also construct matrices such that $\circ(A, B, C)$ exists but neither $A \circ B$ nor $B \circ C$ exists.

We now turn to the question of cascade subtraction. Given matrices A and C, we wish to solve the equation $A \circ X = C$. We assume that all three matrices are partitioned as in (1). We will also use the abbreviations $L = A_{11} - C_{11}$, and $D = A_{22} + X_{11}$.

THEOREM 21. *There exists a solution to the cascade subtraction problem $A \circ X = C$ if and only if all of the following conditions hold:*

(i) $\text{Range}(L) \subset \text{Range}(A_{12})$,

(ii) $\text{Range}(L^*) \subset \text{Range}(A_{21}^*)$,

(iii) $\text{nullity}(A_{12}) \geqq \text{rank}(A_{21}) - \text{rank}(L)$,

(iv) $\text{nullity}(A_{21}^*) \geqq \text{rank}(A_{12}^*) - \text{rank}(L^*)$,

(v) $\text{Range}(C_{12}) \subset \text{Range}(A_{12})$,

(vi) $\text{Range}(C_{21}^*) \subset \text{Range}(A_{21}^*)$.

Proof. First, let us use formula (6) to rewrite $A \circ X = C$ as

$$\begin{aligned} &A_{11} - A_{12}D^- A_{21} = C_{11}, \quad \text{equivalently } L = A_{12}D^- A_{21}, \\ &A_{12}D^- X_{12} = C_{12}, \\ &X_{21}D^- A_{21} = C_{21}, \\ &X_{22} - X_{21}D^- X_{12} = C_{22}. \end{aligned} \tag{8}$$

By the definition of L, we must have

$$\mathscr{S}\begin{bmatrix} L & A_{12} \\ A_{21} & D \end{bmatrix} = 0$$

By Theorem 7 conditions (i)-(iv) above are necessary for the existence of D, and hence X_{11}. Conditions (v) and (vi) are clear from (8).

Conversely, let us assume that (i)-(vi) hold. From (i)-(iv) we know that an X_{11} exists. Then, since $\text{Range}(A_{12}^*) \subset \text{Range}(D^*)$, we have $\text{Range}(A_{12}D^-) = \text{Range}(A_{12})$. In view of (v) we can then take for X_{12} any value of $(A_{12}D^-)^- C_{12}$. Similarly, we can use (vi) to solve for X_{21}. Then $X_{22}=C_{22}+X_{21}D^{-}X_{12}$ and the matrix X has been constructed. Q.E.D.

5. The cascade sum of positive semidefinite matrices. A *resistive network* is an n-port network composed of interconnected resistors. The impedance matrix of such a network will be a real positive semidefinite matrix, although not all such matrices correspond to networks. Thus we are led to consider the cascade sum of positive semidefinite matrices.

Throughout this section we will use the Moore-Penrose generalized inverse, rather than an arbitrary 1-inverse. We do this because the Moore-Penrose generalized inverse of a positive semidefinite matrix will again be positive semidefinite. In fact most of our formulas will not depend on this specific choice of a 1-inverse.

THEOREM 22. *Let A, B and C be positive semidefinite matrices, partitioned in a manner appropriate for forming the cascade sum, and with A and B square. Then*

(i) *A and B are cascade summable,*

(ii) *$A \circ B$ is positive semidefinite,*

(iii) *If $B \leqq C$, then $A \circ B \leqq A \circ C$,*

(iv) $(A \circ B) \circ C = A \circ (B \circ C)$.

Proof. The cascade sum of A and B will be formed by shorting the matrix Q as defined in (5). Since A and B are positive semidefinite, and Q is formed from the direct sum of A and B by a congruence, Q is again positive semidefinite. By Lemma 9 Q is shortable, and thus $A \circ B$ exists and is positive semidefinite. Since congruence preserves partial order, and by Lemma 10 so does shorting, the cascade sum will also preserve partial order. Since the cascade sum always exists, by Theorem 19 the cascade sum will be associative. Q.E.D.

We now wish to discuss the cascade subtraction problem, that is, to solve the equation $A \circ X = C$. Here A and C are positive semidefinite and we seek a positive semidefinite X. As in the previous section, let $D = A_{22}+X_{11}$, and $L = A_{11}-C_{11}$. We also let S be the subspace corresponding to the first block in the partitions for A and C. In terms of the partitioned matrices, the equations are

$$\begin{array}{lll} & \text{(i)} & A_{12}D^{+}A_{21} = L, \\ (9) & \text{(ii)} & A_{12}D^{+}X_{12} = C_{12}, \\ & \text{(iii)} & X_{22}-X_{21}D^{+}X_{12} = C_{22}. \end{array}$$

THEOREM 23. *Let A and C be positive semidefinite matrices. The following conditions are necessary and sufficient for the existence of a positve semidefinite matrix X satisfying $A \circ X = C$.*

(i) *L is a positive semidefinite matrix,*

(ii) $\mathscr{S}(A) \leqq \mathscr{S}(C)$,

(iii) Range (L) = Range (A_{12}),

(iv) Range $(C_{12}) \subset$ Range (A_{12}).

Proof. First, we establish that the conditions are necessary. For (i), observe that X_{11} and A_{22} are both positive semidefinite, and thus $A_{12}D^{+}A_{21}$ must be also. For (ii) observe that $A \circ 0 = \mathscr{S}(A)$. Since cascade addition is monotone it follows that $A \circ 0 \leqq A \circ X = C$. Moreover, Range $(A \circ 0) \subset S$; thus $A \circ 0 \leqq \mathscr{S}(C)$ by Lemma 9 part (a). For (iii) use (i) and Theorem 13. For (iv) we merely use (9)(ii).

In order to see that the conditions are sufficient, recall again that Range $(A_{21}) \subset$ Range (A_{22}). Let $\tilde{A}_{22}$ be the result of shorting A_{22} to the subspace Range (A_{21}); then Range $(\tilde{A}_{22})$ = Range $(A_{22}) \cap$ Range (A_{21}) = Range (A_{21}). By Lemma 3 part (c),

$$A_{12}\tilde{A}_{22}^{+}A_{21} = A_{12}A_{22}^{+}A_{21},$$

and thus

$$\tilde{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & \tilde{A}_{22} \end{bmatrix}$$

is a positive semidefinite operator, and

$$\mathcal{S}(\tilde{A}) = \mathcal{S}(A).$$

The solution of the cascade subtraction problem is then

$$X = \begin{bmatrix} A_{21}L^+A_{12} - \tilde{A}_{22} & A_{21}L^+C_{12} \\ C_{21}L^+A_{12} & C_{22} + C_{21}L^+C_{12} \end{bmatrix}. \tag{10}$$

We must show that X satisfies (9), and that X is positive semidefinite.

First, let us show that (9) is satisfied. By hypothesis we know that L is positive semidefinite with $\operatorname{Range}(L) = \operatorname{Range}(A_{12})$. Therefore by Lemma 3 part (a) $\operatorname{Range}(A_{21}L^+A_{12}) = \operatorname{Range}(A_{21})$. Since by Lemma 9 part (c) $\operatorname{Range}(A_{22} - \tilde{A}_{22}) \cap \operatorname{Range}(A_{21}) = 0$, we have, using Lemma 3 part (c) and then Lemma 3 part (b),

$$A_{12}(A_{21}L^+A_{12} + A_{22} - \tilde{A}_{22})^+A_{21} = A_{12}(A_{21}L^+A_{12})^+A_{21} = A_{12}A_{12}^+LA_{21}^+A_{21} = L.$$

For (9)(i), we have

$$A_{12}D^+A_{21} = A_{12}(A_{21}L^+A_{12} + A_{22} - \tilde{A}_{22})^+A_{21} = L.$$

For (9)(ii), we have

$$A_{12}D^+X_{12} = A_{12}(A_{21}L^+A_{12} + A_{22} - \tilde{A}_{22})^+A_{21}L^+C_{12} = LL^+C_{12} = C_{12}.$$

Finally, for (9)(iii)

$$\begin{aligned} X_{22} - X_{21}D^+X_{12} &= C_{22} + C_{21}L^+C_{12} - C_{21}L^+A_{12}(A_{21}L^+A_{12} + A_{22} - \tilde{A}_{22}) + A_{21}L^+C_{12} \\ &= C_{22} + C_{21}L^+C_{12} - C_{21}L^+LL^+C_{12} = C_{22}. \end{aligned}$$

It remains to prove that X is positive semidefinite. As was done with A_{22}, let $\tilde{C}_{22}$ result from shorting C_{22} to $\operatorname{Range}(C_{21})$. Then $C_{22} \geqq \tilde{C}_{22}$ and thus

$$X \geqq \begin{bmatrix} A_{21}L^+A_{12} - \tilde{A}_{22} & A_{21}L^+C_{12} \\ C_{21}L^+A_{12} & \tilde{C}_{22} + C_{21}L^+C_{12} \end{bmatrix}. \tag{11}$$

We merely need show that the right-hand side of (11) is positive semidefinite. This matrix can be factored as follows

$$\begin{bmatrix} A_{21} & 0 \\ 0 & C_{21} \end{bmatrix} \begin{bmatrix} L^+ - A_{21}^+\tilde{A}_{22}A_{12}^+ & L^+ \\ L^+ & L^+ + C_{21}^+\tilde{C}_{22}C_{12}^+ \end{bmatrix} \begin{bmatrix} A_{12} & 0 \\ 0 & C_{12} \end{bmatrix}.$$

Clearly $L^+ + C_{21}^+\tilde{C}_{22}C_{12}^+$ is positive semidefinite. By Lemma 11 it is thus sufficient to show that $L^+ - A_{21}^+\tilde{A}_{22}A_{12}^+ \geqq L^+(L^+ + C_{21}^+\tilde{C}_{22}C_{12}^+)^+L^+$. Equivalently, using the parallel sum the condition may be rewritten,

$$L^+ : C_{21}^+\tilde{C}_{22}C_{12}^+ \geqq A_{21}^+\tilde{A}_{22}A_{12}^+. \tag{12}$$

Let us assume for the moment that $\operatorname{Range}(C_{12}) = \operatorname{Range}(A_{12})$. Then all terms in (12) have the same range, and we can invert the inequality obtaining

$$L + C_{12}\tilde{C}_{22}^+C_{21} \leqq A_{12}\tilde{A}_{22}^+A_{21}$$

which, in view of the fact that $C_{12}\tilde{C}_{22}^+C_{21} = C_{12}C_{22}^+C_{21}$ is equivalent to hypothesis (ii).

It remains to remove the assumption that $\operatorname{Range}(C_{12}) = \operatorname{Range}(A_{12})$. Let

$$C_\varepsilon = C + \varepsilon \begin{bmatrix} A_{12}A_{22}^+A_{21} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

We will show that for sufficiently small $\varepsilon > 0$ the perturbed system $A \circ X_\varepsilon = C_\varepsilon$ has a positive semidefinite solution and that these solutions tend to X. Thus X is positive semidefinite.

First, let us check the hypotheses of the theorem for C_ε. For (i), we have $L_\varepsilon = A_{11} - (C_{11} + \varepsilon A_{12}A_{22}^{+}A_{21}) = L - \varepsilon A_{12}A_{22}^{+}A_{21}$. Since L is positive semidefinite, and Range $(L) =$ Range $(A_{12}A_{22}^{+}A_{21})$, we will have $L \geqq \varepsilon A_{12}A_{22}^{+}A_{21}$ for sufficiently small ε.

For (ii), note that by Lemma 11 the perturbation term is positive semidefinite, and thus $\mathcal{S}(C_\varepsilon) \geqq \mathcal{S}(C) \geqq \mathcal{S}(A)$.

For (iii), when ε is smaller than the largest ε which makes L_ε positive semidefinite, then Range $(L_\varepsilon) =$ Range $(L) =$ Range (A_{12}).

Finally, for (iv) since Range $(C_{12}) \subset$ Range (A_{12}), Range $(C_{12} + \varepsilon A_{12}) \subset$ Range (A_{12}), and equality holds except for finitely many ε.

Therefore, for sufficiently small $\varepsilon > 0$ the equation $A \circ X_\varepsilon = C_\varepsilon$ has a positive semidefinite solution X_ε, which may be obtained by appropriately modifying (10). In order to establish convergence of the X_ε, it suffices to observe that L_ε^{+} converges to L^{+} since the ranges are the same for sufficiently small ε. Q.E.D.

6. Duality and the chain matrix. The *dual* of a network connection is obtained by interchanging the roles of currents and voltages, see [5] and [23]. For example, the series and parallel network connections are duals.

Prior to writing the equations for the dual cascade connection let us recall that in our original formulation (1) the matrix A is $m_1 + m_2$ by $n_1 + n_2$, and the matrix B is $m_2 + m_3$ by $n_2 + n_3$. Then A_{22} and B_{11} are both m_2 by n_2 so that $D = A_{22} + B_{11}$ is defined. If A and B are cascade summable, then $C = A \circ B$ is $m_1 + m_3$ by $n_1 + n_3$. In the *dual cascade connection* the row and column dimensions are interchanged. That is, the matrix $\textcircled{A}$ is $n_1 + n_2$ by $m_1 + m_2$ and the matrix $\textcircled{B}$ is $n_2 + n_3$ by $m_2 + m_3$.

Given matrices $\textcircled{A}$ and $\textcircled{B}$ and a vector γ we seek vectors α and β such that

$$(13)\quad \begin{array}{ll} \text{(i)} & \left.\begin{array}{l} \alpha_1 = \gamma_1, \\ \beta_2 = \gamma_2, \\ \alpha_2 = \beta_1, \end{array}\right\} \text{(voltage conditions),} \\ \text{(ii)} & \\ \text{(iii)} & \\ \text{(iv)} & A_{21}\alpha_1 + A_{22}\alpha_2 + B_{11}\beta_1 + B_{12}\beta_2 = 0, \quad \text{(current conditions).} \end{array}$$

Given vectors α and β satisfying (13), we then define the current vectors a and b by $a = \textcircled{A}\alpha$ and $b = \textcircled{B}\beta$. The vector c is then defined by $c_1 = a_1$, and $c_2 = b_2$. We note that (13)(iv) could also have been written $a_2 + b_1 = 0$.

DEFINITION. If for all vectors γ there exist vectors α and β satisfying (13), and if moreover, the vector c is uniquely determined by γ, then we say the matrices $\textcircled{A}$ and $\textcircled{B}$ are *dual cascade summable.* As before there will then be a matrix $\textcircled{C}$ such that $\textcircled{C}\gamma = c$. This matrix $\textcircled{C}$ is called the *dual cascade sum* of $\textcircled{A}$ and $\textcircled{B}$; we write $\textcircled{C} = \textcircled{A} \circ' \textcircled{B}$.

Analogous to the construction of the matrix Q in (5), let us define a matrix Q', where $\textcircled{D} = \textcircled{A}_{22} + \textcircled{B}_{11}$.

$$Q' = \begin{bmatrix} \textcircled{A}_{11} & 0 & \textcircled{A}_{12} \\ 0 & \textcircled{B}_{22} & \textcircled{B}_{21} \\ \textcircled{A}_{21} & \textcircled{B}_{12} & \textcircled{D} \end{bmatrix}.$$

Then $\textcircled{A} \circ' \textcircled{B}$ exists if and only if Q' is complementable, yielding the range conditions Range $[\textcircled{A}_{21}, \textcircled{B}_{12}] \subset$ Range $(\textcircled{D})$ and Range $[\textcircled{A}_{12}^{*}, \textcircled{B}_{21}^{*}] \subset$ Range $(\textcircled{D}^{*})$.

THEOREM 24. *If the matrices A and B are cascade summable, then the matrices A^{*} and B^{*} are dual cascade summable.*

Proof. Letting $\textcircled{A} = A^*$ and $\textcircled{B} = B^*$, we have $\text{Range}\,[\textcircled{A}_{21}, \textcircled{B}_{12}] = \text{Range}\,[A_{12}^*, B_{21}^*] = \text{Range}\,[A_{12}^*, -B_{21}^*]$, and $\text{Range}\,(\textcircled{D}) = \text{Range}\,(D^*)$. Therefore $\text{Range}\,[\textcircled{A}_{21}, \textcircled{B}_{12}] \subset \text{Range}\,(\textcircled{D})$ if and only if $\text{Range}\,[A_{12}^*, -B_{21}^*] \subset \text{Range}\,(D^*)$. A similar computation holds for the other range condition. Q.E.D.

We note that if

$$J = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix},$$

then $A^* \circ' B^* = J(A \circ B)^* J$.

THEOREM 25. *Let A and B be cascade summable and invertible. Then if A^{-1} and B^{-1} are dual cascade summable, $A \circ B$ is invertible and $(A \circ B)^{-1} = A^{-1} \circ' B^{-1}$.*

Proof. Let $(A \circ B)c = \gamma$, and let $\textcircled{A} = A^{-1}$ and $\textcircled{B} = B^{-1}$. Then the vectors a, b, α, and β which satisfy (4) also satisfy (13). Since in (13) c is uniquely determined by γ we have $(A^{-1} \circ' B^{-1})(A \circ B)c = (A^{-1} \circ' B^{-1})\gamma = c$, and thus $A^{-1} \circ' B^{-1}$ is a left inverse for $A \circ B$. Proceeding in the other direction, we find that $A \circ B$ is a left inverse for $A^{-1} \circ' B^{-1}$. Therefore $A \circ B$ is invertible, as desired. Q.E.D.

We note that A and B, being invertible, must of course be square. We have not explicitly assumed that the partition is such that $A \circ B$ is square; instead, this fact is part of the conclusion of the theorem. We also note that one can construct 2 by 2 invertible matrices A and B such that the cascade sum exists and is not invertible; hence A^{-1} and B^{-1} are not dual cascade summable.

COROLLARY 26. *If the matrices A and B are positive definite, with A_{22} square, then $A \circ B$ is positive definite, and $(A \circ B)^{-1} = A^{-1} \circ' B^{-1}$.*

Proof. The result follows immediately from the two previous theorems. Q.E.D.

In electrical network theory one standard approach to the cascade sum is by means of the *chain matrix*, see [10], [16], [23]. Next we develop some of the theory of the chain matrix. We will show that the chain matrix approach is not useful for computing the cascade sum except in the special case where A and B are square. Even if one restricts attention to the electrically reasonable situation where A, B, and the submatrix $D = A_{11} + B_{22}$ are square, the chain matrix is still not always appropriate.

For a matrix A, partitioned as in (1), the chain matrix $K(A)$ is defined by

$$(14) \qquad K(A)\begin{bmatrix} \alpha_2 \\ -a_2 \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ a_1 \end{bmatrix} \quad \text{whenever } A\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}.$$

We note that if some vectors $[{}^{\alpha_2}_{-a_2}]$ do not arise in $Aa = \alpha$, then the chain matrix is not defined by (14) for these vectors.

LEMMA 27. *A chain matrix $K(A)$ exists for the matrix A if and only if A_{21} possesses a left inverse. In this case the matrix*

$$(15) \qquad \begin{bmatrix} A_{11}A_{21}^{-L} & A_{11}A_{21}^{-L}A_{22} - A_{12} \\ A_{21}^{-L} & A_{21}^{-L}A_{22} \end{bmatrix}$$

is a chain matrix for A for any choice of the left inverse A_{21}^{-L}.

Proof. If A_{21}^{-L} exists, then we can solve $A_{21}a_1 + A_{22}a_2 = \alpha_2$ for a_1 to obtain (15). Conversely, suppose $A_{21}x = 0$ for a vector x. Then from

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} x \\ 0 \end{bmatrix} = \begin{bmatrix} A_{11}x \\ 0 \end{bmatrix} \quad \text{we have } K(A)\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} A_{11}x \\ x \end{bmatrix}.$$

Therefore $x = 0$. Thus $\text{Kernel}\,(A_{21}) = 0$, so that a left inverse A_{21}^{-L} exists. Q.E.D.

LEMMA 28. *Let A_{21} be invertible. Then $K(A)$ is unique; $K(A)$ possesses a chain matrix $K(K(A))$ and $K(K(A)) = A$.*

Proof. If A_{21} is invertible, then the equation $A_{21}a_1+A_{22}a_2=\alpha_2$ uniquely determines a_1 in terms of a_2 and α_2. Thus all pairs α_2, a_2 are possible, and a_1 and hence α_1 are uniquely determined by α_2 and a_2. Thus $K(A)$ is unique. Since $(K(A))_{21}=A_{21}^{-1}$, which again is invertible, $K(K(A))$ exists. It is immediate from (14) that $K(K(A))=A$. Q.E.D.

We note that if A_{21} is not square, then it is not possible for both $K(A)$ and $K(K(A))$ to be defined. If a nonsquare A_{21} possesses a left inverse then the chain matrix exists but is not unique; moreover, whenever a chain matrix exists it will correspond to many matrices.

THEOREM 29. *Let A and B be cascade summable matrices such that chain matrices $K(A)$ and $K(B)$ exist. Then for any choices of the chain matrices $K(A)$ and $K(B)$ the product $K(A)K(B)$ is a chain matrix for the cascade sum $A\circ B$.*

Proof. Given the current vector c, we let a, b, α, β, and γ be as in (4). Then

$$\begin{bmatrix}\gamma_1\\ c_1\end{bmatrix}=\begin{bmatrix}\alpha_1\\ a_1\end{bmatrix}=K(A)\begin{bmatrix}\alpha_2\\ -a_2\end{bmatrix}=K(A)\begin{bmatrix}\beta_1\\ b_1\end{bmatrix}$$

$$=K(A)K(B)\begin{bmatrix}\beta_2\\ -b_2\end{bmatrix}=K(A)K(B)\begin{bmatrix}\gamma_2\\ -c_2\end{bmatrix}. \qquad \text{Q.E.D.}$$

Even though we have now found a chain matrix for C, we cannot necessarily recover the matrix C; stronger hypotheses are necessary.

THEOREM 30. *Let A and B be matrices with A_{21} and B_{21} invertible. Let $L=K(A)K(B)$ be the product of the chain matrices. Then L_{21} is invertible if and only if A and B are cascade summable, in which case $L=K(A\circ B)$, and $C=K(L)$.*

Proof. First assume that $A\circ B$ exists. In view of the previous theorem we need only show that C_{21} is invertible. If A_{21} is an r by r matrix and B_{21} is s by s, then $D=A_{22}+B_{11}$ is r by s. By cascade summability we have Range $(A_{21})\subset$ Range (D), and therefore rank $(D)\geqq r$. Since also Range $(B_{21}^*)\subset$ Range (D^*), we have rank $(D)\geqq s$. Thus rank $(D)\geqq\max(r,s)$, and hence $r=s=\text{rank}(D)$. Therefore from (6) we have $C_{21}^{-1}=A_{21}^{-1}DB_{21}^{-1}$.

Conversely, using (15) and matrix multiplication to compute L, we have $L_{21}=A_{21}^{-1}(A_{22}+B_{11})B_{21}^{-1}$. Since L_{21} is invertible, the matrix $D=A_{22}+B_{11}$ must be also. Then the range conditions of Theorem 15 are satisfied. By the preceding theorems L is the chain matrix of C, and $C=K(L)$. Q.E.D.

The theorem allows us to present an alternate proof for a special case of Theorem 19. This special case seems to be the only one proved in electrical network literature.

COROLLARY 31. *Let A, B, and C be matrices such that A_{21}, B_{21}, and C_{21} are invertible, and both $(A\circ B)\circ C$ and $A\circ(B\circ C)$ are defined. Then $(A\circ B)\circ C=A\circ(B\circ C)$.*

Proof. The two triple cascade sums are expressed as products of the three chain matrices, and matrix multiplication is associative. Q.E.D.

7. Almost right definite matrices. The real world model that we used to motivate the cascade sum of positive semidefinite matrices was the cascade connection of resistive n-port networks. If the electrical networks contain capacitors or inductors, the impedance matrix will contain *positive real* functions of the frequency, see [11], [17], [23]. In this section we consider one fixed frequency so that the matrices are constant. The appropriate class of matrices is that of the almost right definite matrices [8].

A complex matrix A is said to be *Almost Right Definite* if Re $\langle Aa, a\rangle\geqq 0$ for all a, and if Re $\langle Aa, a\rangle=0$ only if $Aa=0$. Here we are using Re for the real part. A positive semidefinite matrix is of course almost right definite. We will also require the following related concept. A matrix A is *Almost Definite* if $\langle Aa, a\rangle = 0$ only if $Aa = 0$.

LEMMA 32. *Let A and B be almost right definite matrices; then*

(a) *A is almost definite,*

(b) *S^*AS is almost right definite for any matrix S,*

(c) $\begin{bmatrix} A & O \\ O & B \end{bmatrix}$ *is almost right definite.*

Proof. For part (a), if $\langle Aa, a\rangle = 0$, then $\operatorname{Re}\langle Aa, a\rangle = 0$ and thus $Aa = 0$. For part (b), we note that $\langle S^*SAa, a\rangle = \langle ASa, Sa\rangle$; since A is almost right definite we know $\operatorname{Re}\langle ASa, Sa\rangle \geqq 0$ and hence $\operatorname{Re}\langle S^*Asa, a\rangle \geqq 0$. Moreover if $\operatorname{Re}\langle S^*ASa, a\rangle = 0$ then $\operatorname{Re}\langle ASa, Sa\rangle = 0$ and $ASa = 0$ which implies $S^*AS = 0$. Part (c) is obvious. Q.E.D.

THEOREM 33. *Let A and B be almost right definite matrices; then A and B are cascade summable and $A \circ B$ is almost right definite.*

Proof. By Lemma 32, the matrix Q of equation (5) is almost right definite whenever A and B are, and hence Q is almost definite. It is shown in [9], [15] and [18] that every almost definite matrix is complementable. Therefore $A \circ B$ exists.

For a given current c, let a and b be solutions to (4); then we may write following string of equalities

$$\begin{aligned}\langle Aa, a\rangle + \langle Bb, b\rangle &= \langle \alpha_1, a_1\rangle + \langle \alpha_2, a_2\rangle + \langle \beta_1, b_1\rangle + \langle \beta_2, b_2\rangle \\ &= \langle \gamma_1, c_1\rangle + \langle \gamma_2, c_2\rangle = \langle A \circ Bc, c\rangle.\end{aligned}$$

In the second line the two middle terms cancel by (4)(iii) and (4)(iv). Since $\operatorname{Re}\langle Aa, a\rangle \geqq 0$ and $\operatorname{Re}\langle Bb, b\rangle \geqq 0$, we see that $\operatorname{Re}\langle A \circ Bc, c\rangle \geqq 0$. Now suppose $\operatorname{Re}\langle A \circ Bc, c\rangle = 0$, then we must have $\operatorname{Re}\langle Aa, a\rangle = 0$ and $\operatorname{Re}\langle Bb, b\rangle = 0$. Hence $Aa = 0$ and $Bb = 0$, but then $\gamma_1 = 0$ and $\gamma_2 = 0$. Since these voltages are unique, $Cc = 0$ and thus C is almost right definite. Q.E.D.

8. Summary and conclusions. In this work we have carefully developed the theory of the cascade sum of matrices. We have presented conditions which assure the existence of the cascade sum and the associativity of the cascade operation. In particular, for the case of positive semidefinite matrices, the cascade sum is always defined and always associative.

We have also discussed the question of cascade subtraction, both in the general case and in the network synthesis related case of positive semidefinite matrices.

Two areas of interest were not considered in this work. Cascade addition is not commutative, even in the positive semidefinite case. Given a matrix A, determining which matrices B cascade commute with A is a difficult problem. A seemingly unrelated problem is the question of the cascade limit, which involves the analysis of the infinite sequence $A, A \circ A, A \circ A \circ A, \cdots$. It turns out that when the sequence has a limit, call it B, that B is a matrix which cascade commutes with the matrix A. Ando [10] has obtained some very interesting results concerning the cascade limit of positive semidefinite matrices. However, there seems to be much work remaining in the general case.

REFERENCES

[1] W. N. ANDERSON, JR., *Shorted operators*, SIAM J. Appl. Math., 20 (1971), pp. 520–525.

[2] W. N. ANDERSON, JR., R. J. DUFFIN AND G. E. TRAPP, *Parallel subtraction of matrices*, Proc. Nat. Acad. Sci. U.S.A., 69 (1972), pp. 2530–2531.

[3] W. N. ANDERSON, D. REYNOLDS AND G. E. TRAPP, *Cascade addition of matrices*, Proc. WV Acad. Sci., 46 (1974), pp. 185–192.

[4] W. N. ANDERSON, JR. AND G. E. TRAPP, *Shorted operators* II, SIAM J. Appl. Math., 28 (1975), pp. 60–71.

[5] W. N. ANDERSON, JR., R. J. DUFFIN AND G. E. TRAPP, *Matrix operations induced by network connections*, SIAM J. Control, 13 (1975), pp. 446–461.

[6] W. N. ANDERSON, JR. AND G. E. TRAPP, *Inequalities for the parallel connection of resistive n-port networks*, J. Franklin Inst., 299 (1975), pp. 305–313.

[7] ———, *Matrix operations induced by electrical network connections—a survey*, in Constructive Approaches to Mathematical Models, Academic Press, New York, 1979, pp. 53–73.

[8] ———, *Analytic operator functions and electrical networks*, in Recent Applications of Generalized Inverses, S. L. Campbell, ed., Pitman, Boston, 1982, pp. 12–26.

[9] T. ANDO, *Generalizied Schur complements*, Linear Algebra Appl., 27 (1979), pp. 173–186.

[10] ———, *Limited of cascade iteration of matrices*, Numer. Funct. Anal. Optim., 2 (1980), pp. 579–589.

[11] V. BELEVITCH, *Classical Network Theory*, Holden-Day, San Francisco, 1968.

[12] D. CARLSON, E. HAYNSWORTH AND T. MARKHAM, *A generalization of the Schur complement by means of the Moore-Penrose inverse*, SIAM J. Appl. Math., 26 (1974), pp. 254–259.

[13] D. CARLSON, *Matrix decompositions involving the Schur complement*, SIAM J. Appl. Math., 28 (1975), pp. 577–587.

[14] ———, *What are Schur complements, anyway*?, Linear Algebra Appl., 59 (1984), pp. 189–193.

[15] D. CARLSON AND E. V. HAYNSWORTH, *Complementable and almost definite matrices*, Linear Algebra Appl., 52/53 (1983), pp. 157–176.

[16] L. COLLET, *Definition et characteristiques de circuits superposes*, Ann. Telecommumn., 4 (1949), pp. 42–56.

[17] R. J. DUFFIN, *Elementary operations which generate network matrices*, Proc. Amer. Math. Soc., 11 (1963), pp. 645–658.

[18] R. J. DUFFIN AND T. D. MORLEY, *Almost definite matrices and electromechanical systems*, SIAM J. Appl. Math., 35 (1978), pp. 21–30.

[19] JON LEE, *Dynamical cascade models for Kolmogorov's inertial flow*, J. Fluid Mech., 101 (1980), pp. 349–376.

[20] S. K. MITRA AND C. R. RAO, *Generalized Inverse of Matrices and Its Applications*, John Wiley, New York, 1971.

[21] S. K. MITRA AND M. L. PURI, *Parallel sum and difference of matrices*, J. Math. Anal. Appl., 44 (1973), pp. 92–97.

[22] ———, *Shorted operators—an extended concept and some applications*, Linear Algebra Appl., 47 (1982), pp. 57–59.

[23] R. W. NEWCOMB, *Linear Multiport Synthesis*, McGraw-Hill, New York, 1966.

[24] A. H. ZEMANIAN, *Infinite networks of positive operators*, Circuit Theory and Appl., 2 (1974), pp. 69–78.

SIAM J. ALG. DISC. METH.
Vol. 7, No. 4, October 1986

013

MATHEMATICAL ASPECTS OF THE RELATIVE GAIN ARRAY ($A \circ A^{-T}$)*

CHARLES R. JOHNSON† AND HELENE M. SHAPIRO‡

Abstract. For nonsingular n-by-n matrices A, we investigate the map

$$A \rightarrow \Phi(A) \equiv A \circ (A^{-1})^T$$

in which $\circ$ denotes the Hadamard (entry-wise) product. The matrix $\Phi(A)$ arises in mathematical control theory in chemical engineering design problems, where it is known as the relative gain array, and also in a matrix theoretic problem involving the relation between the diagonal entries and eigenvalues. We first give several elementary properties of Φ and show that the iterates $\Phi^k(A)$ converge to I when A is either positive definite or an H-matrix. We then discuss, with examples and partial results, several unsolved problems associated with Φ. These include the range of Φ, inverse images of elements in the range of Φ, fixed points of Φ, etc.

Key words. relative gain array, Hadamard product

AMS(MOS) subject classifications. 15, 93

1. Introduction. The *Hadamard* (or entry-wise) *product* of two $n \times m$ matrices $B = (b_{ij})$ and $C = (c_{ij})$ is defined by

$$B \circ C = (b_{ij}c_{ij}).$$

For nonsingular $n \times n$ matrices A, we investigate the map

$$A \rightarrow \Phi(A) \equiv A \circ A^{-T},$$

where "A^{-T}" means the inverse transpose, $(A^{-1})^T$, of A. When we write this, we presume the inverse of A exists.

The product $A \circ A^{-T}$ has arisen in two distinct contexts and may be of interest in a third.

In mathematical control theory associated with chemical engineering design problems, $\Phi(A)$ is known as the *relative gain array* (RGA) associated with the gain matrix A. Bristol introduced the RGA in [1]; it has since enjoyed a variety of applications [6]. A typical problem is to control a number of outputs $x_1, \cdots, x_n$ by manipulating certain inputs $m_1, \cdots, m_n$. For example, the outputs might be the products of a distillation process in which the rates of flow of the inputs are controlled by valves. The gain matrix A is the Jacobian of the x's with respect to the m's, and the RGA is used in a static analysis of a proposed plant configuration. Generally, changing one input will affect several outputs. However, for simplicity, one would like to primarily control each output by manipulating a single input, and it is desirable to pair the inputs and outputs with this in mind. The RGA has been used to find preferred pairings. If σ is a permutation of $\{1, 2, \cdots, n\}$, we call the entries $a_{1\sigma(1)}, a_{2\sigma(2)}, \cdots, a_{n\sigma(n)}$ a *diagonal* of the matrix $A = (a_{ij})$. In certain chemical engineering applications a diagonal of the RGA in which the entries are "near" 1 is used to determine the pairing of inputs

* Received by the editors May 13, 1985, and in revised form January 29, 1986.

† Department of Mathematical Sciences, Clemson University, Clemson, South Carolina 29631. The work of this author was supported in part by National Science Foundation grants MCS 80-01611 and DMS-8500372 and by Office of Naval Research contract N00014-86-K-0012.

‡ Department of Mathematics, Swarthmore College, Swarthmore, Pennsylvania 19081. The work of this author was carried out while visiting the Institute for Physical Science and Technology, College Park, Maryland 20742.

and outputs for further design analysis. In most applications thus far, the matrices are small ($n=2, 3$, or 4) and the "good" diagonal is chosen by inspection of the RGA. Though more sophisticated mathematical analysis of the RGA has begun, our purpose here is to contribute to the general understanding of the mapping Φ. In particular, one possibility for determining a permutation σ, and thus a pairing, may arise from analysis of iteration of Φ.

The matrix $\Phi(A)$ also appears in the following matrix theoretic problem. Suppose $\lambda_1, \cdots, \lambda_n$ and $b_{11}, \cdots, b_{nn}$ are two sets of n numbers. Mirsky [8] showed that there exists an $n \times n$ matrix B with eigenvalues $\lambda_1, \cdots, \lambda_n$ and diagonal entries $b_{11}, \cdots, b_{nn}$ if and only if $b_{11}+\cdots+b_{nn}=\lambda_1+\cdots+\lambda_n$. Now, *if B is diagonalizable*, we have $B=ADA^{-1}$, where $D=\operatorname{diag}(\lambda_1, \cdots, \lambda_n)$ and A is nonsingular. Writing $A=(a_{ij})$ and $A^{-1}=(\alpha_{ij})$, we see that

$$b_{ii}=\sum_{j=1}^{n} a_{ij}\lambda_j\alpha_{ji},$$

and thus

$$(*) \qquad (b_{11}, \cdots, b_{nn})^T=\Phi(A)(\lambda_1, \cdots, \lambda_n)^T.$$

In § 5 we prove a stronger version of Mirsky's result and use this to show that, except in the case $\lambda_1=\cdots=\lambda_n$, a necessary and sufficient condition for there to exist a nonsingular A such that (*) holds is that $b_{11}+\cdots+b_{nn}=\lambda_1+\cdots+\lambda_n$.

We mention a third, more speculative, potential application of Φ, associated with computational evidence that iterates Φ^k(A) often converge to a permutation matrix associated with a "large" diagonal of A. In the travelling salesman's problem, a minimum distance tour of given "cities" is sought. If the i, j entry of A is inversely related to the distance between city i and city j, and if $\Phi(A)$ converges to a permutation, this permutation may be an approximate salesman's tour and iteration of Φ may provide a travelling salesman heuristic worthy of study.

This paper is divided into two parts. The first part includes some elementary observations about the map Φ and features two convergence results about iterates $\Phi^k(A)$. In particular, we show that

$$\lim_{k\to\infty} \Phi^k(A)=I$$

if either A is a positive definite matrix or an H-matrix. In the second part we discuss a number of interesting open mathematical problems concerning the map Φ and include some relevant fragmentary results and examples. These questions include: what is the range of Φ; what are the fixed points; when does $\Phi(B)=\Phi(A)$; for which A do iterates of Φ converge and to what, etc.?

PART I

2. Elementary properties. We begin with some elementary properties of $A \circ A^{-T}$ and the map Φ.

Observation 1. The matrix $A \circ A^{-T}$ has row and column sums 1.

Observation 2. If D and E are $n \times n$, nonsingular, diagonal matrices, then

$$\Phi(DAE)=\Phi(A).$$

Observation 3. $\Phi(A^{-1})=[\Phi(A)]^T=\Phi(A^T)$.

Observation 4. If P and Q are $n \times n$ permutation matrices, then

$$\Phi(PAQ)=P\Phi(A)Q.$$

Observation 5. If P is an $n\times n$ permutation matrix, then

$$\Phi(P)=P.$$

Observation 6. If A is a nonsingular triangular matrix, then

$$\Phi(A)=I.$$

Observation 7. If A is reducible (i.e., if PAP^T is block upper or lower triangular for some permutation P) then $\Phi(A)$ is completely reducible and $P\Phi(A)P^T$ is a direct sum of diagonal blocks corresponding to the diagonal blocks of PAP^T.

Observation 1 follows from the Laplace expansion for $\det(A)$ in terms of the cofactors of a fixed row or column. Observation 2 follows from $D(A\circ B)=(DA)\circ B=A\circ(DB)$ and $(A\circ B)E=A\circ(BE)=(AE)\circ B$. To obtain 3, use $A\circ B=B\circ A$ and $(A\circ B)^T=A^T\circ B^T$. Numbers 4 and 5 follow from $P^{-1}=P^T$ and $(PAQ)\circ(PBQ)=P(A\circ B)Q$. Six holds because the inverse of an upper (lower) triangular matrix is again upper (lower) triangular, and the diagonal entries of the inverse are the inverses of the diagonal entries of the original matrix. Similarly, the inverse of a block upper (lower) triangular matrix is block upper (lower) triangular with diagonal blocks which are the inverses of the diagonal blocks in the original matrix. Observation 7 then follows from this fact and from Observation 4. However, the example below shows that $\Phi(A)$ may be reducible when A is irreducible.

Example 1. Let

$$A=\begin{pmatrix}1 & 1 & 1\\ \frac{1}{2} & 1 & 0\\ 1 & 2 & 1\end{pmatrix}.$$

Then A is irreducible. However,

$$\Phi(A)=\begin{pmatrix}2 & -1 & 0\\ 1 & 0 & 0\\ -2 & 2 & 1\end{pmatrix}$$

is reducible.

3. Convergence of $\Phi^k(A)$ for positive definite A. Throughout this section, A will be a positive definite Hermitian matrix. The term "positive definite" will mean "positive definite Hermitian". In order to deal with the complex case, we define $\Phi_*(A)=A\circ A^{-*}$. When A is real, $\Phi(A)=\Phi_*(A)$; when A is Hermitian, $\Phi_*(A)=A\circ A^{-1}$. A theorem of Schur states that when A and B are positive definite, $A\circ B$ is positive definite [9], [10]. Hence $\Phi_*(A)=A\circ A^{-1}$ is positive definite. Using induction, one sees that $\Phi_*^k(A)$ is positive definite for every positive integer k. Thus, we may define the sequence $A^{(k)}=\Phi_*^k(A)$, and each $A^{(k)}$ is positive definite. We put $A=\Phi^0(A)=A^{(0)}$, and denote the ij entry of $A^{(k)}$ as $a_{ij}^{(k)}$. Our goal is the following theorem.

THEOREM 1. *Let A be a positive definite Hermitian matrix. Then* $\lim_{k\to\infty}\Phi_*^k(A)=I$.

The proof relies on some known results and some preliminary lemmas.

If A and B are Hermitian, then $A\geqq B$ means $A-B$ is positive semidefinite.

THEOREM A. 1. (Fiedler) [2], [7]. *If A is a positive definite Hermitian matrix, then* $A\circ A^{-T}\geqq I$.

2. (Johnson) [5], [11]. *If A is a positive definite Hermitian matrix then* $A\circ A^{-1}\geqq I$.

Theorem A tells us that $A^{(k)}\geqq I$ for all $k\geqq 1$.

For a Hermitian matrix H, let $\lambda_1(H)\leqq\lambda_2(H)\leqq\cdots\leqq\lambda_n(H)$ denote the eigenvalues of H in increasing order.

THEOREM B (Schur) [9], [10]. *If A and B are positive semidefinite, then*

$$\lambda_1(A)\min\{b_{11},\cdots,b_{nn}\}\leqq\lambda_j(A\circ B)\leqq\lambda_n(A)\max\{b_{11},\cdots,b_{nn}\}.$$

We use Theorem B to establish Lemma 1.

LEMMA 1. *Let A be positive definite, with $A\neq I$ and $\lambda_1(A)\geqq 1$. Then*

$$\lambda_n(A\circ A^{-1})<\lambda_n(A).$$

Proof. By Theorem B,

$$\lambda_n(A\circ A^{-1})\leqq\lambda_n(A^{-1})\max\{a_{11},\cdots,a_{nn}\}.$$

Now $\lambda_n(A^{-1})=1/\lambda_1(A)\leqq 1$ and $a_{ii}\leqq\lambda_n(A)$ for each i. Furthermore, since $\lambda_n(A)$ is the largest eigenvalue of A, $a_{ii}=\lambda_n(A)$ only if $\lambda_n(A)$ is the only nonzero entry in row i and column i. Suppose exactly r diagonal entries of A are equal to $\lambda_n(A)$. If $r=0$, then $\max\{a_{11},\cdots,a_{m}\}<\lambda_n(A)$ and we are done. If $r=n$, then $A=\lambda_n(A)I$ and $A\circ A^{-1}=I$. Since $A\neq I$, we have $\lambda_n(A\circ A^{-1})=1<\lambda_n(A)$. If $0<r<n$, we may apply a permutation similarity to A and thus assume $A=\lambda_n(A)I_k\oplus A_1$, where A_1 is $n-k\times n-k$, positive definite, $\lambda_1(A_1)\geqq 1$, and $\lambda_n(A_1)\leqq\lambda_n(A)$. Now $A\circ A^{-1}=I_k\oplus(A_1\circ A_1^{-1})$, so $\lambda_n(A\circ A^{-1})=\lambda_n(A_1\circ A_1^{-1})$. Since $\lambda_n(A)$ is not a diagonal entry of A_1, we have $\lambda_n(A_1\circ A_1^{-1})<\lambda_n(A)$. Hence, $\lambda_n(A\circ A^{-1})<\lambda_n(A)$. □

LEMMA 2. *Let A be positive definite and $A^{(k)}=\Phi_*^{(k)}(A)$. Then either $A^{(k)}=I$ for all sufficiently large k, or the sequence $\lambda_n(A^{(k)})$, for $k\geqq 2$, is strictly decreasing and bounded below by* 1.

Proof. Theorem A tells us $A^{(k)}\geqq I$ for $k\geqq 1$, so $\lambda_n(A^{(k)})\geqq 1$ when $k\geqq 1$. The rest of the statement follows immediately from Lemma 1. □

Recall that the quantity $\|M\|=\max_{x\neq 0}|x^*Mx/x^*x|$ defines a norm on the set of $n\times n$ complex matrices, known as the spectral norm. When A is positive definite, $\|A\|=\lambda_n(A)$. Lemma 2 then tells us that the sequence $A^{(k)}$ either reaches I in a finite number of steps, or else is monotonically decreasing in norm for $k\geqq 2$. Thus, the sequence $\{A^{(k)}\}$ is bounded in norm. We are now ready to prove Theorem 1.

Proof of Theorem 1. If $A^{(r)}=I$ for some r, then $A^{(k)}=I$ for all $k\geqq r$, and we are done. So, assume $A^{(k)}\neq I$ for any k. The sequence $\{A^{(k)}\}$ is bounded in norm; hence it has a convergent subsequence. Let

$$\mathscr{L}=\{L|L\text{ is the limit of a convergent subsequence of }\{A^{(k)}\}\}.$$

Since $A^{(k)}$ is positive definite and $A^{(k)}\geqq I$ for all k, we must have L positive definite and $L\geqq I$ for all $L\in\mathscr{L}$. Now the sequence $\{\lambda_n(A^{(k)})\}$ is strictly decreasing and bounded below by 1, so $\lambda=\lim_{k\to\infty}\lambda_n(A^{(k)})$ exists and $\lambda\geqq 1$. Furthermore, $\lambda=\|L\|=\lambda_n(L)$ for every $L\in\mathscr{L}$.

We now claim $\Phi_*(\mathscr{L})\subseteq\mathscr{L}$. Let $L\in\mathscr{L}$. Then $L=\lim_{j\to\infty}A^{(k_j)}$, where $k_1<k_2<k_3<\cdots$ is an increasing sequence of positive integers. The map Φ_* is continuous, so

$$\begin{aligned}\Phi_*(L)&=\Phi_*\left(\lim_{j\to\infty}\Phi_*(A^{k_j})\right)\\&=\lim_{j\to\infty}\Phi_*(\Phi_*^{k_j}(A))\\&=\lim_{j\to\infty}\Phi_*^{k_j+1}(A)=\lim_{j\to\infty}A^{k_j+1}.\end{aligned}$$

Hence $\Phi_*(L)\in\mathscr{L}$. Now each $L\in\mathscr{L}$ satisfies $\lambda_1(L)\geqq 1$, so by Lemma 1, either $L=I$ or $\lambda_n(\Phi_*(L))<\lambda_n(L)$. But $\Phi(L)\in\mathscr{L}$ implies $\lambda_n(\phi_*(L))=\lambda=\lambda_n(L)$. Therefore, $L=I$ and

$\mathcal{L}=\{I\}$. This means every convergent subsequence of $\{A^{(k)}\}$ converges to I. Therefore, $\lim_{k\to\infty} A^{(k)}=I$. □

Remarks. If one could show directly that $\lim_{k\to\infty}\lambda_n(A^{(k)})=1$, this might yield a cleaner proof of Theorem 1. Some information about the rate of decrease of the sequence $\lambda_n(A^{(k)})$ would be helpful. Also, note that the results and proofs hold for the map Φ as well as Φ_*.

Finally, we mention a result which is not strong enough to establish Theorem 1, but seems interesting in its own right. Perhaps a stronger version of the proposition below would give some insight into the rate of convergence of $A^{(k)}$.

PROPOSITION 1. *Let A be positive definite Hermitian and let $B=A\circ A^{-1}$. Then*

$$\left|\frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}}\right| \leqq \left|\frac{a_{ij}}{\sqrt{a_{ii}a_{jj}}}\right|$$

with equality if and only if $a_{ij}=0$.

Proof. A and B are both positive definite so $a_{ii}>0$ and $b_{ii}>0$ for $i=1,\cdots,n$. Put

$$D_A=\operatorname{diag}\left(\frac{1}{\sqrt{a_{11}}},\cdots,\frac{1}{\sqrt{a_{nn}}}\right) \quad\text{and}\quad D_B=\operatorname{diag}\left(\frac{1}{\sqrt{b_{11}}},\cdots,\frac{1}{\sqrt{b_{nn}}}\right).$$

Let $A_1=D_AAD_A$ and $B_1=D_BBD_B$. Both A_1 and B_1 have ones on the main diagonal. The off diagonal entries of A_1 and B_1 are, respectively,

$$\frac{a_{ij}}{\sqrt{a_{ii}a_{jj}}} \quad\text{and}\quad \frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}}.$$

Now $B=A\circ A^{-1}=A_1\circ A_1^{-1}$. Thus,

$$B_1=D_BBD_B=D_B(A_1\circ A_1^{-1})D_B=A_1\circ(D_BA_1^{-1}D_B).$$

Let α_{ij} denote the i,j entry of $D_BA_1^{-1}D_B$. Since A_1 and B_1 have 1's on the diagonal, $\alpha_{ii}=1$ for $i=1,\cdots,n$. Then $D_BA_1^{-1}D_B$ positive definite implies $|\alpha_{ij}|<1$ for $i\neq j$. Hence, $B_1=A_1\circ(D_BA_1^{-1}D_B)$ tells us

$$\left|\frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}}\right| = \left|\frac{a_{ij}}{\sqrt{a_{ii}a_{jj}}}\right||\alpha_{ij}| \leqq \left|\frac{a_{ij}}{\sqrt{a_{ii}a_{jj}}}\right|$$

with equality if and only if $a_{ij}=0$. □

4. Convergence of $\Phi^k(A)$ when A is an H-matrix. Let A be an $n\times n$ complex matrix. Define

$$R_i(A)=\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}| \quad\text{and}\quad r_i(A)=\frac{R_i(A)}{|a_{ii}|},$$

assuming $a_{ii}\neq 0$. We say A is *row diagonally dominant* provided $R_i(A)<|a_{ii}|$ (or equivalently, $r_i(A)<1$) for each $i=1,\cdots,n$. We say A is an *H-matrix* if there exists a diagonal matrix D such that AD is diagonally dominant. An equivalent definition is the following. Define the comparison matrix $C=C(A)$ of A by putting $c_{ii}=|a_{ii}|$ for $i=1,\cdots,n$ and $c_{ij}=-|a_{ij}|$ when $i\neq j$. Then A is an H matrix if and only if $C(A)$ is an M-matrix. In this section we prove:

THEOREM 2. *If A is an H-matrix, then* $\lim_{k\to\infty}\Phi_k(A)=I$.

Since $\Phi(A)=\Phi(AD)$, we may assume that A itself is row diagonally dominant. We break the argument into several lemmas. The main idea is that $\Phi(A)$ is "more" diagonally dominant than A, hence $\Phi^k(A)$ converges to a diagonal matrix as $k\to\infty$.

LEMMA 3. *Let A be row diagonally dominant and let B be the row echelon form of A, obtained by doing Gaussian elimination (without row exchanges) on A. Then B is also row diagonally dominant and*

$$r_i(B)\leqq r_i(A).$$

Proof. It suffices to consider the effect of an elementary row operation; one may then apply induction. At the first stage of Gaussian elimination, we clear out column 1 of A by subtracting (a_{j1}/a_{11}) (Row 1 of A) from row j of A. Row 1 of A is not changed; the new row j, for $j\geqq 2$, is:

$$0 \quad a_{j2}-\left(\frac{a_{j1}}{a_{11}}\right)a_{12} \quad a_{j3}-\left(\frac{a_{j1}}{a_{11}}\right)a_{13} \quad \cdots \quad a_{jn}-\left(\frac{a_{j1}}{a_{11}}\right)a_{1n}.$$

The new jth diagonal entry is

$$a_{jj}-\left(\frac{a_{j1}}{a_{11}}\right)a_{1j}.$$

Now

$$\left|\frac{a_{j1}}{a_{11}}\right||a_{1j}|=|a_{j1}|\left|\frac{a_{1j}}{a_{11}}\right|<|a_{j1}|<|a_{jj}|,$$

so

$$\begin{aligned}
r_j(A)\left|a_{jj}-\left(\frac{a_{j1}}{a_{11}}\right)a_{1j}\right| &\geqq r_j(A)|a_{jj}|-r_j(A)\left|\frac{a_{j1}}{a_{11}}\right||a_{1j}|\\
&=\sum_{\substack{k=1\\k\neq j}}^{n}|a_{jk}|-r_j(A)\left|\frac{a_{j1}}{a_{11}}\right||a_{1j}|\\
&=\sum_{\substack{k=2\\k\neq j}}^{n}|a_{jk}|+\left|\frac{a_{j1}}{a_{11}}\right|(|a_{11}|-r_j(A)|a_{1j}|)\\
&\geqq\sum_{\substack{k=2\\k\neq j}}^{n}|a_{jk}|+\left|\frac{a_{j1}}{a_{11}}\right|\left(\sum_{k=2}^{n}|a_{1k}|-|a_{1j}|\right)\\
&=\sum_{\substack{k=2\\k\neq j}}^{n}\left(|a_{jk}|+\left|\frac{a_{j1}}{a_{11}}\right||a_{1k}|\right)\\
&\geqq\sum_{\substack{k=2\\k\neq j}}^{n}\left|a_{jk}-\left(\frac{a_{j1}}{a_{11}}\right)a_{1k}\right|.
\end{aligned}$$

Thus, we see that the typical elementary row operation yields a matrix which is at least as diagonally dominant as the original matrix. □

We use A_{ij} to denote the $n-1\times n-1$ minor obtained by striking out row i and row j of A.

LEMMA 4. *If A is row diagonally dominant, then*

$$|\det A_{ij}| \leqq r_j(A)|\det A_{ii}|, \quad \textit{for } i \neq j.$$

Proof. One can easily verify that this holds for 2×2 matrices. We assume the inequality holds for $n-1\times n-1$ matrices and use induction. Let A be $n\times n$ and diagonally dominant. Applying a simultaneous permutation of rows and columns, we may move a_{ii} into the n, n position and a_{ij} into the $n, 1$ position. Since this permutation similarity does not affect diagonal dominance, it suffices to show

$$|\det A_{n1}| \leqq r_1(A)|\det A_{nn}|.$$

Gaussian elimination does not change $\det A_{ni}$ for any $i = 1, \cdots, n$, and by Lemma 1, the ratios $r_i(A)$ cannot increase. Hence, we may assume A is in row echelon form. Thus,

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ & a_{22} & a_{23} & \cdots & a_{2n} \\ & & a_{33} & \cdots & a_{3n} \\ & & & & \vdots \\ & & & & a_{nn} \end{pmatrix}$$

is upper triangular. We have $|\det A_{nn}| = |a_{11}||a_{22}| \cdots |a_{n-1\,n-1}|$. Let $A[1, n|1, j]$ denote the $n-2\times n-2$ minor obtained by removing rows 1 and n and columns 1 and j from A. Expanding $\det A_{n1}$ by cofactors along the first row yields

$$\det A_{n1} = \sum_{j=2}^{n} a_{1j}(-1)^j \det A[1, n|1, j].$$

Notice that $A[1, n|1, j]$ is obtained by removing row n and column j from A_{11}. Since A_{11} is of size $n-1$, and is diagonally dominant, the induction hypothesis tells us

$$|\det A[1, n|1, j]| \leqq r_j(A)|\det A[1, n|1, n]|.$$

Hence

$$|\det A[1, n|1, j]| \leqq r_j(A)|a_{22}||a_{33}| \cdots |a_{n-1\,n-1}|.$$

Since $r_j(A) < 1$ for each j, we see

$$\begin{aligned} |\det A_{n1}| &\leqq |a_{22}||a_{33}| \cdots |a_{n-1\,n-1}|\left(\sum_{j=2}^{n} |a_{1j}|\right) \\ &= |a_{11}||a_{22}| \cdots |a_{n-1\,n-1}|r_1(A). \end{aligned}$$

Thus, $|\det A_{n1}| \leqq r_1(A)|\det A_{nn}|$. □

Remark. Lemma 2 tells us that if A is row diagonally dominant, then A^{-1} is diagonally dominant in its column entries—i.e., each diagonal entry of A^{-1} dominates the other entries in that column.

LEMMA 5. *Let A be row diagonally dominant and put $\gamma(A) = \max\{r_1(A), r_2(A), \cdots, r_n(A)\}$. Then $\Phi(A) = A \circ A^{-T}$ is diagonally dominant and*

$$\gamma(\Phi(A)) \leqq (\gamma(A))^2.$$

Proof. The i, j entry of $A \circ A^{-T}$ is

$$\frac{(-1)^{i+j}a_{ij}(\det A_{ij})}{\det A}.$$

Let $B = A \circ A^{-T}$. Then

$$\begin{aligned}\sum_{\substack{j=1\\ j\neq i}}^{n} |b_{ij}| &= \frac{1}{|\det A|}\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}||\det A_{ij}|\\ &\leqq \frac{1}{|\det A|}\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|r_j(A)|\det A_{ii}|\\ &\leqq \frac{\gamma(A)}{|\det A|}|\det A_{ii}|\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\\ &\leqq \gamma(A)(r_i(A)|a_{ii}|)\left|\frac{\det A_{ii}}{\det A}\right|\\ &\leqq (\gamma(A))^2|b_{ii}|. \qquad \square\end{aligned}$$

We now prove Theorem 2.

Proof of Theorem 2. Let A be an H matrix. Then AD is row diagonally dominant for some diagonal matrix D. Since $\Phi(A) = \Phi(AD)$, we may assume A itself is diagonally dominant. Since diagonal dominance implies nonsingularity, Lemma 5 guarantees that $\Phi^k(A)$ is row diagonally dominant and hence nonsingular for every k.

Let $A^{(k)} = \Phi^k(A)$ and set $\gamma = \gamma(A)$. Then $0 \leqq \gamma < 1$ and by Lemma 5, $\gamma(A^{(k)}) \leqq \gamma^{(2^k)}$. Now let D_k be the diagonal matrix

$$\operatorname{diag}\left(\frac{1}{a_{ii}^{(k)}}, \cdots, \frac{1}{a_{nn}^{(k)}}\right),$$

then $D_kA^{(k)}$ has ones on the main diagonal. Note that D_k merely rescales the rows of $A^{(k)}$, so $D_kA^{(k)}$ is still diagonally dominant and $\gamma(D_kA^{(k)}) = \gamma(A^{(k)})$. Put $B^{(k)} = D_kA^{(k)}$; then $b_{ii}^{(k)} = 1$ for $i = 1, \cdots, n$ and $\gamma(B^{(k)}) \leqq \gamma^{(2^k)}$. We have $A^{(k+1)} = \Phi(A^{(k)}) = \Phi(D_kA^{(k)}) = \Phi(B^{(k)}) = B^{(k)} \circ (B^{(k)})^{-T}$. Let $\beta_{ij}^{(k)}$ be the ij entry of $(B^{(k)})^{-T}$. Then $\beta_{ij}^{(k)} = (-1)^{i+j}(\det B_{ij}^{(k)}/\det B^{(k)})$. Since $B^{(k)}$ has ones on the main diagonal, while $|b_{ij}^{(k)}| \leqq \gamma^{2^k}$ for $i \neq j$, we see that $\det B^{(k)} = 1 + \varepsilon_k$, where $\lim_{k\to\infty} |\varepsilon_k| = 0$. Similarly, $\det B_{ii}^{(k)} = 1 + \delta_k$ where $\lim_{k\to\infty} |\delta_k| = 0$. Hence, $\lim_{k\to\infty} \beta_{ii}^{(k)} = 1$. Therefore, $\lim_{k\to\infty} a_{ii}^{(k+1)} = 1$. Since $\lim_{k\to\infty} \gamma(A^{(k)}) \leqq \lim_{k\to\infty} \gamma^{2^k} = 0$, we have $\lim_{k\to\infty} a_{ij}^{(k+1)} = 0$. Hence $\lim_{k\to\infty} A^{(k)} = I$. □

PART 2

5. The range of Φ and diagonals of similar matrices. A basic question about the map Φ is to determine its range, both for the case where the domain consists of real, nonsingular matrices, and for the domain of complex, nonsingular matrices. We know that every matrix in the range of Φ has row and column sums 1. Is the converse true, or are there additional conditions which further restrict the range of Φ? The following example shows that not every matrix with row and column sums 1 is the image of a real matrix, under Φ. The problem for the complex case is not resolved.

Example 2. Let $\zeta_n = e^{2\pi i/n}$ and let V_n denote the $n \times n$ Vandermonde matrix with $\zeta^{(i-1)(j-1)}$ in position i, j. Then $V_n^{-1} = V_n^{-T} = (1/n)V_n^* = (1/n)\bar{V}_n$. Thus, $\Phi(V_n) = (1/n)V_n \circ \bar{V}_n = (1/n)J_n$, where J_n denotes the $n \times n$, all one matrix. However, direct calculation shows that the equation $\Phi(A) = \frac{1}{3}J_3$ has no real solutions.

Example 2 raises the question: when does $\Phi(A) = (1/n)J_n$ have a real solution? If H_n is an $n \times n$ Hadamard matrix, then $\Phi(H_n) = (1/n)J_n$. However, a necessary condition for a Hadamard matrix of size n to exist is that 4 divide n; it is not known if this necessary condition is also sufficient.

Recall from the introduction, that when $D = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$ and $B = A^{-1}DA$, we have

$$(*) \qquad (b_{11}, \cdots, b_{nn})^T = \Phi(A)(\lambda_1, \cdots, \lambda_n)^T.$$

We will show that, except when $\lambda_1 = \lambda_2 = \cdots = \lambda_n$, a necessary and sufficient condition for (*) to hold for some nonsingular A is that

$$b_{11} + b_{22} + \cdots + b_{nn} = \lambda_1 + \lambda_2 + \cdots + \lambda_n.$$

This follows immediately from the following result about diagonals of similar matrices.

THEOREM 3. *Let A be an $n \times n$ matrix over a field F, with either* $\operatorname{char}(F) = 0$ *or* $\operatorname{char}(F) \geqq n$, *and suppose $A \neq \alpha I$, i.e., A is not a scalar matrix. Let $b_{11}, b_{22}, \cdots, b_{nn} \in F$. Then*

$$b_{11} + b_{22} + \cdots + b_{nn} = \operatorname{tr}(A)$$

if and only if $b_{11}, b_{22}, \cdots, b_{nn}$ are the diagonal entries of an $n \times n$ matrix B which is similar to A and has entries from F.

To prove Theorem 3, we use an induction argument which requires that we first establish the results for $n = 2$. Also, the cases $n = 3$ and $n = 4$ involve some special arguments which we isolate in separate lemmas. The fact that similar matrices have the same trace establishes the "if" part of Theorem 3, so we need only prove the "only if" half.

LEMMA 6. *Theorem* 3 *holds for $n = 2$.*

Proof. Let A be a 2×2, nonscalar matrix. If A is not diagonal, at least one of the entries a_{12}, a_{21} is nonzero. Assume $a_{12} \neq 0$; the other case is similar.

Let

$$x = \frac{b_{11} - a_{11}}{a_{12}}$$

and define

$$S = \begin{pmatrix} 1 & 0 \\ x & 1 \end{pmatrix}.$$

Note that $b_{11} = a_{12}x + a_{11}$ and $b_{22} = a_{11} + a_{22} - b_{11} = a_{22} - a_{12}x$. Direct computation then shows that $B = S^{-1}AS$ has diagonal entries b_{11} and b_{22}.

If A is diagonal, we have

$$A = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}$$

with $\lambda \neq \mu$. Put

$$r = \frac{b_{11} - \mu}{\lambda - \mu};$$

then

$$b_{11} = r\lambda + (1 - r)\mu$$

and

$$b_{22} = \lambda + \mu - b_{11} = (1 - r)\lambda + r\mu.$$

The matrix

$$S=\begin{pmatrix} r & 1 \\ 1-r & -1 \end{pmatrix}$$

is nonsingular and $B=S^{-1}AS$ has diagonal entries b_{11} and b_{22}. □

LEMMA 7. *Theorem* 3 *holds for* $n=3$.

Proof. Let A be a nonscalar, 3×3 matrix with entries in F. Then A has a 2×2 principal submatrix which is not scalar; we may assume, without loss of generality, that

$$A_{12}=\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

is not scalar. As shown in Lemma 6, for each i, we can find a nonsingular matrix R_i such that $R_i^{-1}A_{12}R_i$ has diagonal entries b_{ii} and $a_{11}+a_{22}-b_{ii}$. Thus, for $T_i=R_i\oplus 1$, with $i=1, 2, 3$, we have

$$T_i^{-1}AT_i=\left(\begin{array}{c|cc} b_{ii} & * & * \\ \hline * & \multicolumn{2}{c}{\multirow{2}{*}{A_i}} \\ * & & \end{array}\right),$$

where A_i is 2×2. If it is possible to choose b_{ii} so that A_i is not a scalar matrix, we may apply Lemma 6 to A_i, and thus obtain the desired B. Note that we can adjust the order of the diagonal entries with a permutation similarity.

If $A_i=\lambda_i I_2$, for $i=1, 2, 3$, further analysis is needed.

If $\lambda_1=\lambda_2=\lambda_3=\lambda$, then $b_{11}=b_{22}=b_{33}=\lambda$, and $T_i^{-1}AT_i$ is the desired B.

If $\lambda_1=\lambda_2=\lambda$, but $\lambda_3\neq\lambda$, then

$$b_{11}=b_{22}=\operatorname{tr}(A)-2\lambda$$

while

$$b_{33}=\operatorname{tr}(A)-2\lambda_3.$$

Then

$$\operatorname{tr}(A)=b_{11}+b_{22}+b_{33}=3\operatorname{tr}(A)-4\lambda-2\lambda_3.$$

Since char $(F)\neq 2$, this yields

$$\operatorname{tr}(A)=2\lambda+\lambda_3.$$

Hence, $b_{11}=b_{22}=\lambda_3$ and $b_{33}=2\lambda-\lambda_3$. The desired B is then $T_3^{-1}AT_3$.

Finally, suppose λ_1, λ_2, λ_3 are all distinct. Each λ_i must be an eigenvalue of A, so A has the three distinct eigenvalues λ_1, λ_2, λ_3 and

$$\operatorname{tr}(A)=\lambda_1+\lambda_2+\lambda_3.$$

Since $b_{ii}+2\lambda_i=\operatorname{tr}(A)$ for each i, we have

$$b_{11}=-\lambda_1+\lambda_2+\lambda_3,$$

$$b_{22}=\lambda_1-\lambda_2+\lambda_3,$$

$$b_{33}=\lambda_1+\lambda_2-\lambda_3.$$

Since $\lambda_i=\frac{1}{2}(\operatorname{tr}(A)-b_{ii})$, the eigenvalues λ_i are in F, and A is similar, via a similarity over F, to

$$D=\operatorname{diag}(\lambda_1, \lambda_2, \lambda_3).$$

Now $\lambda_1 \neq \lambda_2$, so diag (λ_1, λ_2) is similar to

$$\begin{pmatrix} b_{11} & * \\ * & \lambda_1+\lambda_2-b_{11} \end{pmatrix}.$$

But $\lambda_1+\lambda_2-b_{11}=2\lambda_1-\lambda_3$, so A is similar to a matrix of the form

$$\begin{pmatrix} b_{11} & * & 0 \\ * & 2\lambda_1-\lambda_3 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}.$$

Since $\lambda_1 \neq \lambda_3$, $2\lambda_1-\lambda_3 \neq \lambda_3$. Also $b_{22}+b_{33}=2\lambda_1$, so

$$\begin{pmatrix} 2\lambda_1-\lambda_3 & 0 \\ 0 & \lambda_3 \end{pmatrix}$$

is similar to a 2×2 matrix with diagonal entries b_{22} and b_{33}. Hence A is similar, over F, to some B with diagonal entries b_{11}, b_{22}, b_{33}. □

The last special case to settle is $n=4$.

LEMMA 8. *Theorem* 3 *holds for* $n=4$.

Proof. Let A be a 4×4, nonscalar matrix over F. As in the proof of Lemma 7, we can find nonsingular matrices T_i, for $i=1,\cdots,n$, such that

$$T_i^{-1}AT_i=\left(\begin{array}{c|ccc} b_{ii} & * & * & * \\ \hline * & & & \\ * & & A_i & \\ * & & & \end{array}\right),$$

where A_i is 3×3. If A_i is nonscalar for some i, we apply Lemma 7 to conclude that there exists a B similar to A with diagonal entries $b_{11},\cdots,b_{44}$.

Otherwise, $A_i=\lambda_i I_3$, for each i. Then each λ_i is an eigenvalue of A of multiplicity at least two. There are then two possibilities. Either $\lambda_1=\lambda_2=\lambda_3=\lambda_4=\lambda$, or there are exactly two distinct λ_i's, say λ_1 and λ_2, and A has eigenvalues λ_1 and λ_2, each of multiplicity 2.

If $\lambda_1=\lambda_2=\lambda_3=\lambda_4=\lambda$, then $b_{ii}=\operatorname{tr}(A)-3\lambda$ for each i. From this, it follows that $b_{ii}=\lambda$ for each i and $B=T_i^{-1}AT_i$ is the desired matrix.

In the remaining case, we have $\lambda_1\neq\lambda_2$ and

$$\operatorname{tr}(A)=2\lambda_1+2\lambda_2.$$

Hence

$$b_{11}=\operatorname{tr}(A)-3\lambda_1=2\lambda_2-\lambda_1$$

and

$$b_{22}=\operatorname{tr}(A)-3\lambda_2=2\lambda_1-\lambda_2.$$

Since $b_{33}+b_{44}=\lambda_1+\lambda_2$, and since the only possible values for any b_{ii} are $2\lambda_2-\lambda_1$ and $2\lambda_1-\lambda_2$, we have (reordering if necessary),

$$b_{11}=b_{33}=2\lambda_2-\lambda_1,$$

$$b_{22}=b_{44}=2\lambda_1-\lambda_2.$$

Since $\lambda_1+\lambda_2=\frac{1}{2}(\operatorname{tr}(A))\in F$ and b_{11}, $b_{22}\in F$, we see $3\lambda_1\in F$ and $3\lambda_2\in F$. Since char $F\neq 3$, the eigenvalues λ_1 and λ_2 are in F. Hence A is similar, over F, to a triangular matrix of the form

$$T=\begin{pmatrix}\lambda_1 & * & * & * \\ 0 & \lambda_2 & * & * \\ 0 & 0 & \lambda_1 & * \\ 0 & 0 & 0 & \lambda_2\end{pmatrix}.$$

Since $b_{11}+b_{22}=b_{33}+b_{44}=\lambda_1+\lambda_2$, we may apply Lemma 6 to the 2×2 blocks in rows and columns 1 and 2, and rows and columns 3 and 4 to obtain a B similar to A with diagonal entries b_{11}, b_{22}, b_{33}, b_{44}. □

We can now present the proof of Theorem 3.

Proof of Theorem 3. We use induction on n. Thus, we assume the result holds for $n-1$ and let A be an $n\times n$, nonscalar matrix. Since we have already proven the result for $n\leqq 4$, we assume $n\geqq 5$. Then A has a nonscalar, 2×2 principal submatrix, and as in the proof of Lemma 7, we can find nonsingular T_i, for $i=1,\cdots,n$, such that

$$T_i^{-1}AT_i=\left(\begin{array}{c|c} b_{ii} & * \cdots * \\ \hline * & \\ \vdots & A_i \\ * & \end{array}\right)$$

where A_i is $n-1\times n-1$. If A_i is not scalar for some i, then the induction hypothesis says there exists a nonsingular R_i, of size $n-1\times n-1$ such that $B_i=R_i^{-1}A_iR_i$ has diagonal entries $b_{11},\cdots,b_{i-1,i-1}$, $b_{i+1,i+1},\cdots,b_{nn}$. Put $S=T_i(1\oplus R_i)$, then $B=S^{-1}AS$ has the required diagonal entries.

If $A_i=\lambda_iI_{n-1}$ for $i=1,\cdots,n$, then each λ_i is an eigenvalue of A of multiplicity at least $n-2$. Since $n\geqq 5$, the matrix A has at most one eigenvalue of multiplicity $n-2$, or greater, hence $\lambda_1=\lambda_2=\cdots=\lambda_n=\lambda$. Then $b_{11}=b_{22}=\cdots=b_{nn}$. Since $b_{ii}=\operatorname{tr}(A)-(n-1)\lambda$, and char $(F)\neq n-1$, we must have $b_{ii}=\lambda$ for every i. Hence $T_i^{-1}AT_i$ is the desired matrix B. □

Remark. Mirsky [8] proved that when $\lambda_1+\cdots+\lambda_n=b_{11}+\cdots+b_{nn}$, there exists a matrix B with eigenvalues $\lambda_1,\cdots,\lambda_n$ and diagonal entries $b_{11},\cdots,b_{nn}$. Theorem 3 is a stronger version of this result—it guarantees the existence of a B which not only has the same eigenvalues as A, but is actually similar to A.

COROLLARY 1. *Suppose* $\lambda_1,\cdots,\lambda_n$ *are not all equal. Then there exists a nonsingular matrix A such that*

$$(b_{11},\cdots,b_{nn})^T=\Phi(A)(\lambda_1,\cdots,\lambda_n)^T$$

if and only if

$$b_{11}+b_{22}+\cdots+b_{nn}=\lambda_1+\lambda_2+\cdots+\lambda_n.$$

Proof. Since $\Phi(A)$ has column sums 1, necessity follows by multiplying both sides of

$$(b_{11},\cdots,b_{nn})^T=\Phi(A)(\lambda_1,\cdots,\lambda_n)^T$$

on the left by the all one matrix J.

To prove sufficiency, let $D=\operatorname{diag}(\lambda_1,\cdots,\lambda_n)$. Then D is not a scalar matrix, and Theorem 3 gives the result. □

Corollary 1 is of interest in the following context. Recall that if $x_1 \leqq x_2 \leqq \cdots \leqq x_n$ and $y_1 \leqq y_2 \leqq \cdots \leqq y_n$ are two sets of real numbers, then there exists a doubly stochastic matrix P such that

$$P(x_1, \cdots, x_n)^T = (y_1, \cdots, y_n)^T$$

if and only if the vector $(y_1, \cdots, y_n)$ majorizes the vector $(x_1, \cdots, x_n)$—i.e., $y_1 \geqq x_1$, $y_1+y_2 \geqq x_1+x_2$, $\cdots y_1+y_2+\cdots+y_i \geqq x_1+x_2+\cdots+x_i, \cdots, y_1+\cdots+y_n = x_1, \cdots, x_n$. Furthermore, it is known [3], [8] that there exists a Hermitian matrix with eigenvalues $\lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_n$ and diagonal entries $h_{11} \leqq h_{22} \leqq \cdots \leqq h_{nn}$ if and only if $(h_{11}, \cdots, h_{nn})$ majorizes $(\lambda_1, \cdots, \lambda_n)$. Since H can be diagonalized by a unitary matrix U, this means that the doubly stochastic matrix which transforms the λ's to the h's can actually be chosen to be the orthostochastic matrix $U \circ U^{-T} = U \circ \bar{U}$. Horn [3] gave the following example to show that not every doubly stochastic matrix is orthostochastic.

Example 3. Let

$$A = \frac{1}{2}\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Then A is doubly stochastic, but $A \neq U \circ \bar{U}$ for any unitary matrix U. This raises analogous questions about the range of Φ. If $\Phi(A)$ is nonnegative, then $\Phi(A)$ must be doubly stochastic. Furthermore, we see that the range of Φ is "large enough" to provide a $\Phi(A)$ to transform $(\lambda_1, \cdots, \lambda_n)$ to $(b_{11}, \cdots, b_{nn})$, (assuming the λ's are not all equal, and $\lambda_1+\cdots\lambda_n = b_n+\cdots+b_{nn}$), just as the set of orthostochastic matrices is "large enough" to transform $x=(x_1, \cdots, x_n)$ into $y=(y_1, \cdots, y_n)$ whenever y majorizes x. Furthermore, the matrix of Example 2 is in the range of Φ; in fact it is its own image under Φ. Does the range of Φ contain all doubly stochastic matrices? If not, is there a nice characterization of those doubly stochastic matrices which are also in the range of Φ?

We conclude this section with one more example. It is known that an $n \times n$, doubly stochastic matrix cannot have an $r \times s$ block of zeros when $r+s \geqq n+1$. The example below shows this is not true for $\Phi(A)$.

Example 4. Let

$$\mathrm{A} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

Then

$$\Phi(A) = \begin{pmatrix} -1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

has a 2×2 block of zeros.

6. Other problems. We briefly mention some other unsolved problems concerning Φ.

Problem 1. Suppose B is in the range of Φ. Is there some nice description of $\{A|\Phi(A)=B\}$, the inverse image of B? Observation 2 tells us that $\Phi(A) = \Phi(\mathrm{DAE})$ for any nonsingular, diagonal D and E. Are there conditions on $\Phi(A)$ which would guarantee that $\Phi(A) = \Phi(C)$ only when $C = DAE$? Notice that the inverse image of the identity matrix includes all triangular matrices, so perhaps some sort of irreducibility condition is needed. The example below shows that $\Phi(A)=I$ can hold even when A is neither triangular nor essentially triangular, i.e., of the form PTP^T, where T is triangular and P is a permutation.

Example 5. Let

$$A=\begin{pmatrix}1&1&0&0\\0&1&1&1\\1&1&1&0\\0&0&1&1\end{pmatrix}.$$

Then $\Phi(A)=I$, but the graph of A shows that A is irreducible and, therefore, not essentially triangular. (This example was suggested by a phone conversation with Ed Bristol.)

Problem 2. In Part 1, we showed that $\lim_{k\to\infty}\Phi^k(A)=I$ when A is positive definite symmetric or when A is an H-matrix. The totally positive matrices (i.e., real matrices for which all of the determinantal minors are positive) are another class of interest. We shall say A is *inverse totally positive* if A^{-1} is totally positive. The matrix A then has the "checkerboard" sign pattern (i.e., a_{ij} has sign $(-1)^{i+j}$) and $\Phi(A)$ also has the checkerboard sign pattern, as does $A\circ A^{-1}$. Is either of $\Phi(A)$ or $A\circ A^{-1}$ inverse totally positive? If $\Phi(A)$ is totally positive, does the sequence of iterates $\Phi^k(A)$ converge to I?

Problem 3. When do the iterates $\Phi^k(A)$ converge, or else terminate with a singular matrix? When does the sequence $\Phi^k(A)$ converge to a permutation matrix? If $B=PAQ$, where P and Q are permutation matrices, we say A and B are permutation equivalent. Since $\Phi(B)=P\Phi(A)Q$, Theorems 1 and 2 tell us that whenever B is permutation equivalent to either an H-matrix, or a positive definite symmetric matrix, the iterates $\Phi^k(B)$ converge to a permutation. Is there some weaker type of diagonal dominance which will guarantee convergence to a permutation?

7. Fixed points of Φ. Suppose $\lim_{k\to\infty}\Phi^k(A)=R$. Then R is a fixed point of Φ—i.e., $\Phi(R)=R$. The most obvious fixed points of Φ are the permutations. However, there are others, and we shall see by an example that fixed points can be more complicated than one might first guess. We do not have a complete description of the fixed points of Φ, but mention some properties and exhibit examples.

PROPOSITION 2. *Let P and Q be permutation matrices. Then*

1) $\Phi(P)=P$.

2) *If $P+Q$ is nonsingular,*

$$\Phi(\tfrac{1}{2}(P+Q))=\tfrac{1}{2}(P+Q).$$

Furthermore, for $0\leqq\alpha\leqq 1$, the iterates $\Phi^k(\alpha P+(1-\alpha)Q)$ converge to P when $\alpha>\frac{1}{2}$ and to Q when $\alpha<\frac{1}{2}$.

Proof. We have already observed (1). Since $\Phi(\frac{1}{2}(P+Q))=P\Phi(\frac{1}{2}(I+P^{-1}Q))$, it suffices to prove (2) for $P=I$.

Let C_r denote the $r\times r$ circulant matrix

$$\begin{pmatrix}0&1&0&\cdots&0\\0&0&1&\cdots&0\\0&0&0&\cdots&1\\1&0&0&\cdots&0\end{pmatrix}.$$

The matrix C_r represents an r-cycle. Since any permutation may be expressed as a product of disjoint cycles, any permutation matrix may be reduced, via a permutation similarity, to a direct sum of cycles. Thus, it suffices to consider $\Phi(I+C_r)$. The matrix $I+C_r$ is nonsingular if and only if r is odd, in which case

$$(I+C_r)^{-1}=\tfrac{1}{2}(I-C_r+C_r^2-C_r^3+\cdots-C_r^{r-2}+C_r^{r-1}).$$

Using $C_r^T=C_r^{-1}=C_r^{r-1}$, one easily computes

$$\Phi(\tfrac{1}{2}(I+C_r))=\tfrac{1}{2}(I\circ I+C_r\circ C_r)=\tfrac{1}{2}(I+C_r).$$

Similarly, computing $\Phi^k(\alpha P+(1-\alpha)Q)$ reduces to the problem of finding $\Phi^k(\alpha I+(1-\alpha)C_r)$. For r odd, and $\beta\neq-1$,

$$(I+\beta C_r)^{-1}=\frac{1}{1+\beta^r}(I-\beta C_r+\beta^2C_r^2-\cdots-\beta^{r-2}C^{r-2}+\beta^{r-1}C^{r-1}).$$

We have

$$\Phi(I+\beta C_r)=\frac{1}{1+\beta^r}(I+\beta^rC_r).$$

For positive integers k,

$$\Phi^k(I+\beta C_r)=\frac{1}{1+\beta^{(r^k)}}(I+\beta^{(r^k)}C_r).$$

For $\alpha\neq0$,

$$\Phi(\alpha I+(1-\alpha)C_r)=\Phi(I+\beta C_r),$$

where $\beta=(1-\alpha)/\alpha$. When $0<\alpha<\frac{1}{2}$, we have $\beta>1$; when $\frac{1}{2}<\alpha\leqq1$, we have $0\leqq\beta<1$. Hence,

$$\lim_{k\to\infty}\Phi^k(\alpha I+(1-\alpha)C_r)$$

is C_r when $0\leqq\alpha<\frac{1}{2}$ and I when $\frac{1}{2}<\alpha\leqq1$. □

Remark. Theorem 2 and Observation 4 tell us that if A is "close" to the permutation P (i.e., $|a_{ij}-p_{ij}|<\varepsilon$ for suitably small ε), then $\{\Phi^k(A)\}$ converges to P. Notice this behavior does not hold near the fixed point $\frac{1}{2}(P+Q)$—the nearby points $\alpha P+(1-\alpha)Q$, where α is close to $\frac{1}{2}$, determine sequences $\Phi^k(\alpha P+(1-\alpha)Q)$ which converge either to P (for $\alpha>\frac{1}{2}$) or Q (for $\alpha<\frac{1}{2}$), but not to $\frac{1}{2}(P+Q)$.

Let J denote the all one matrix. Direct computation shows that if J is $n\times n$, the matrix $(1/(n-1))(J-I)$ is a fixed point of Φ. This is a special case of a more general result.

PROPOSITION 3. *The matrix A is a fixed point of Φ if and only if $(J-A^{-T})/(n-1)$ is a fixed point.*

Proof. The proof relies on the following formula. Let A be an $n\times n$ invertible matrix. Let u and v be $n\times1$ column vectors and suppose $1+u^TA^{-T}v\neq0$. Then

$$(A+uv^T)^{-1}=A^{-1}-\frac{(A^{-1}u)(A^{-T}v)^T}{1+u^TA^{-T}v}.$$

For A nonsingular, define

$$f(A)=\frac{J-A^{-T}}{n-1}.$$

Let e denote the all one vector; then $J=ee^T$. Suppose $\Phi(A)=A$. Then A has row and column sums 1; hence $Ae=A^Te=e$ and $e^TA=e^TA^T=e^T$. Using the formula above,

$$(A^{-1}-J)^{-1}=A+\frac{ee^T}{1-e^Te}=A+\frac{J}{1-n}.$$

Thus,

$$\begin{aligned}\Phi(f(A))&=\Phi(J-A^{-T})=(J-A^{-T})\circ(J-A^{-T})^{-T}\\&=(J-A^{-T})\circ(J-A^{-1})^{-1}\\&=(A^{-T}-J)\circ\left(A+\frac{J}{1-n}\right)\\&=A^{-T}\circ A-A+\frac{A^{-T}}{1-n}-\frac{J}{1-n}.\end{aligned}$$

Since A is a fixed point, $A^{-T}\circ A=A$ and $\Phi(f(A))=(A^{-T}-J)/(1-n)=f(A)$. Now, when A is a fixed point, $f^2(A)=A$. Hence, A is a fixed point if and only if $f(A)=(J-A^{-T})/(n-1)$ is a fixed point, and the map f acts as an involution on the set of fixed points. □

Remark. Since $\Phi(A^T)=(\Phi(A))^T$, we see that A is a fixed point if and only if A^T is a fixed point. Thus, if any one of the four matrices A, A^T, $(J-A^{-T})/(n-1)$, $(J-A^{-1})/(n-1)$ is a fixed point, so are the other three.

We collect together some elementary properties of fixed points in Proposition 4.

PROPOSITION 4. *Let $A=(a_{ij})$ be a fixed point and let $A^{-T}=(\alpha_{ij})$. Then A has the following properties.*

1) *A has row and column sums* 1.
   A^{-T} has row and column sums 1.
2) *If $a_{ji}\neq 0$, then $\alpha_{ij}=1$, or* $\det A_{ij}=(-1)^{i+j}(\det A)$.
3) *Every row and column of A has at least one zero entry.*
4) *At least one of the two matrices A and $(J-A^{-T})/(n-1)$ has at least $n^2/2$ zeros.*
5) *If A is permutation equivalent to a block triangular matrix $\left(\begin{smallmatrix}R & S\\ 0 & T\end{smallmatrix}\right)$, with R of size k, T of size $n-k$, and an $n-k\times k$ block of zeros,* 0, *then $S=0$, and so A is permutation equivalent to $R\oplus S$.*
6) *Let $\mathcal{R}_i$ denote the subset of $\{1,\cdots,n\}$ corresponding to the positions of the nonzero entries in row i of A. Then $\mathcal{R}_i\not\subseteq\mathcal{R}_j$ for any $i\neq j$. The same holds for columns of A.*

Proof. 1) and 2) follow immediately from $A\circ A^{-T}=A$.

If $a_{ij}\neq 0$, then $\alpha_{ij}=1$ and the ij entry of $J-A^{-T}$ is zero. This observation, together with Proposition 3, yields statements (3) and (4).

Statement (5) is apparent from

$$\begin{pmatrix}R & S\\ 0 & T\end{pmatrix}^{-1}=\begin{pmatrix}R^{-1} & -R^{-1}ST^{-1}\\ 0 & T^{-1}\end{pmatrix}.$$

To establish statement (6), suppose $\mathcal{R}_i\subseteq\mathcal{R}_j$ for some $i\neq j$. Without loss of generality, assume $\mathcal{R}_1\subseteq\mathcal{R}_2$. By reordering the columns of A, we may assume $\mathcal{R}_1=\{1,2,3,\cdots,k\}$ and $\mathcal{R}_2=\{1,2,\cdots,k,k+1,\cdots,k+r\}$. But then (2) tells us row 1 of A^{-T} has 1's in positions 1 through k, while row 2 of A^{-T} has 1's in positions 1 through $k+r$. Since A has row sums 1, this forces the inner product of row 1 of A with row 2 of A^{-T} to be 1, which is impossible. Hence $\mathcal{R}_i\not\subseteq\mathcal{R}_j$.

The observations made in Proposition 4 enable us to completely determine the fixed points of Φ when $n \leqq 4$.

PROPOSITION 5. 1) *The only* 2×2 *fixed points are*

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \textit{and} \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

2) *Every* 3×3 *fixed point is either a permutation or an average of two permutations,* $\frac{1}{2}(P+Q)$, *where* $P+Q$ *is invertible.*

3) *If* A *is a real* 4×4 *fixed point, then either* A *or* $(J-A^{-T})/(n-1)$ *is permutation equivalent to a direct sum of fixed points of smaller order. If* A *is not real, but is a* 4×4 *fixed point, then either* A *or* $(J-A^{-T})/(n-1)$ *is permutation equivalent to the matrix*

$$\left(\frac{1}{1+\omega}\right)(I+\omega C_4)=\left(\frac{1}{1+\omega}\right)\begin{pmatrix} 1 & \omega & 0 & 0 \\ 0 & 1 & \omega & 0 \\ 0 & 0 & 1 & \omega \\ \omega & 0 & 0 & 1 \end{pmatrix}$$

or $(1/(1-\bar{\omega}))(I+\bar{\omega}C_4)$, *where* $\omega = e^{\pi i/3}$.

Proof. The case $n=2$ follows immediately from statement (3) of Proposition 4 and Observation 2.

Suppose $n=3$ and A is a fixed point which is not a permutation. A cannot be permutation equivalent to a direct sum, for such a matrix would be a permutation. Therefore, every row and column of A has exactly two nonzero entries. Without loss of generality, we assume the zeros are on the main diagonal. Then $(J-A^{-T})/2$ is a diagonal matrix; hence $(J-A^{-T})/2=I$. Thus,

$$A^{-T}=J-2I \quad \text{and} \quad A=\frac{1}{2}\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Now let $n=4$. Suppose neither A nor $(J-A^{-T})/3$ is permutation equivalent to a direct sum. Then every row and column of A must have exactly two zero entries. Permutating rows and columns of A, we may assume every diagonal entry is nonzero. Then, after simultaneous row and column permutations (which preserve the main diagonal) and using the fact that A is not equivalent to a direct sum, we have

$$A=\begin{pmatrix} \alpha & 1-\alpha & 0 & 0 \\ 0 & \alpha & 1-\alpha & 0 \\ 0 & 0 & \alpha & 1-\alpha \\ 1-\alpha & 0 & 0 & \alpha \end{pmatrix}.$$

Thus, $A=\alpha I+(1-\alpha)C_4$. Now when $|\alpha| \neq |1-\alpha|$, we know the sequence $\Phi^k(A)$ converges to I when $|\alpha|>|1-\alpha|$ and to C_4 when $|1-\alpha|>|\alpha|$. If $|\alpha|=|1-\alpha|$, and α is real, $\alpha=\frac{1}{2}$. But $I+C_4$ is singular. Hence, if A is a real fixed point, either A or $(J-A^{-T})/3$ must be a permutation equivalent to a direct sum.

If we permit complex values for α, then $|\alpha|=|1-\alpha|$ implies $1-\alpha=e^{i\theta}\alpha$ or $\alpha=1/(1+e^{i\theta})$ for some $0\leqq\theta\leqq 2\pi$. Note $\theta\neq\pi$. Then $A=(1/(1+e^{i\theta}))(I+e^{i\theta}C_4)$. Now, for $e^{4i\theta}\neq 1$, the matrix $I+e^{i\theta}C_4$ is nonsingular, and

$$(I+e^{i\theta}C_4)^{-1}=\frac{1}{1-e^{4i\theta}}(I-e^{i\theta}C_4+e^{2i\theta}C_4^2-e^{3i\theta}C_4^3).$$

Thus,

$$\Phi(A)=\Phi(I+e^{i\theta}C_4)=\left(\frac{1}{1-e^{4i\theta}}\right)(I-e^{4i\theta}C_4).$$

To have $\Phi(A)=A$, we must have $e^{i\theta}=-e^{4i\theta}$. Hence $e^{3i\theta}=-1$, and $e^{i\theta}$ is a primitive sixth root of 1. Thus, $\theta=\pi/3$ or $-\pi/3$. Let $\omega=e^{(\pi/3)i}$. Then $\omega^4=-\omega$ and

$$\Phi(I+\omega C_4)=\frac{1}{1+\omega}(I+\omega C_4).$$

Thus, $(1/(1+\omega))(I+\omega C_4)$ is a fixed point, as is $(1/(1+\bar{\omega}))(I+\bar{\omega}C_4)$. □

Thus, we see that for $n\leqq 4$, the fixed points are severely restricted; in fact there are only a finite number of them. Furthermore, when $n\leqq 4$, the real fixed points are nonnegative and hence doubly stochastic. The example below shows that when $n\geqq 5$, there are infinitely many fixed points, and the real ones need not be stochastic.

Example 6. Let a be any real or complex number. Put

$$A=\frac{1}{2}\begin{pmatrix}1 & 1 & 0 & 0 & 0\\ 1 & 0 & 1 & 0 & 0\\ 0 & a & 1-a & 1 & 0\\ 0 & 1-a & a & 0 & 1\\ 0 & 0 & 0 & 1 & 1\end{pmatrix}.$$

Then

$$A^{-T}=\begin{pmatrix}1 & 1 & -1 & 1-2a & 2a-1\\ 1 & -1 & 1 & 2a-1 & 1-2a\\ -1 & 1 & 1 & 1 & -1\\ -1 & 1 & 1 & -1 & 1\\ 1 & -1 & -1 & 1 & 1\end{pmatrix}.$$

The matrix A satisfies $\Phi(A)=A$, regardless of the value of a. Note that if $a=0$ or $a=1$, A is an average of two permutations.

REFERENCES

[1] E. Bristol, *On a new measure of interaction for multivariable process control*, IEEE Trans. Automat. Control, AC-11 (1961), p. 133.

[2] M. Fiedler, *Über eine Ungleichung fur positiv definite Matrizen*, Math. Nachr., 23 (1961), pp. 197-199.

[3] A. Horn, *Doubly stochastic matrices and the diagonal of a rotation matrix*, Amer. J. Math., 76 (1954), pp. 620-630.

[4] C. Johnson, *Hadamard products of matrices*, Linear and Multilinear Algebra, 1 (1974), pp. 295-307.

[5] ———, *Partitioned and Hadamard product matrix inequalities*, J. Res. Nat. Bur. Standards, 83 (1978), pp. 585-591.

[6] Thomas J. McAvoy, *Interaction Analysis: Principles and Applications*, Instrument Society of America, 1983.

[7] T. Markham, *An application of theorems of Schur and Albert*, Proc. Amer. Math. Soc., 59 (1976), pp. 205-210.

[8] L. Mirsky, *Matrices with prescribed characteristic roots and diagonal elements*, J. London Math. Soc., 33 (1958), pp. 14-21.

[9] I. Schur, *Bemerkungen zur theorie der beschrankten bilinearformen mit unendlich vielen veranderlichen*, J. Reine Angew. Math., 140 (1911), pp. 1-28.

[10] G. Styan, *Hadamard products and multivariate statistical analysis*, Linear Algebra Appl., 6 (1973), pp. 217-240.

[11] C. R. Johnson and L. Elsher, *The relationship between Hadamard and conventional multiplication for positive definite matrices*, Linear Algebra Appl., to appear.