SIAM J. ALG. DISC. METH.
Vol. 3, No. 1, March 1982

0196-5212/82/0301-0001 $01.00/0

THE ALGEBRAIC RICCATI EQUATION WITHOUT COMPLETE CONTROLLABILITY*

H. K. WIMMER†

Abstract. The equation $XDX + XA + A^*X - C = 0$ is studied under the assumptions that D is positive semidefinite and to pure imaginary eigenvalues of the Hamiltonian matrix $\left(\begin{smallmatrix} A & D \\ C & -A^* \end{smallmatrix}\right)$ correspond only elementary divisors of even degree. A necessary and sufficient condition on the existence and uniqueness of hermitian solutions is given. The approach is based on symplectic transformations of Hamiltonian matrices.

1. Introduction.

The algebraic Riccati equation

$$XDX + XA + A^*X - C = 0, \tag{1.1}$$

where X, A, C and D are complex $n \times n$ matrices and C and D and the unknown matrix X are hermitian, has a wide range of applications and has been studied extensively. Surveys on (1.1) can be found in the papers by Willems [12], Kučera [6], Coppel [2] and Molinari [10]. The algebraic investigation on existence and uniqueness of solutions in [2] is based on the use of symplectic transformations. In this note we refine Coppel's method and extend one of his results ([2, Thm. 6]). Our assumption will be $D \geqq 0$, i.e. D is positive semidefinite. We do not require the pair (A, D) to be controllable.

2. Symplectic matrices and solutions.

The matrix

$$M = \begin{pmatrix} A & D \\ C & -A^* \end{pmatrix} \tag{2.1}$$

is closely related to (1.1). Obviously X is a solution of (1.1) if and only if

$$(-X \quad I) M \begin{pmatrix} I \\ X \end{pmatrix} = 0 \tag{2.2}$$

holds. With respect to the $2n \times 2n$ matrix J,

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix},$$

M is hermitian, i.e.,

$$JM = (JM)^* \tag{2.3}$$

and M is called a *Hamiltonian* matrix. A matrix $R \in \mathbb{C}^{2n \times 2n}$ is said to be *symplectic* if

$$RJR^* = J \tag{2.4}$$

holds. Under a symplectic similarity property (2.3) is preserved, $JR^{-1}MR = (JR^{-1}MR)^*$. If R is partitioned into $n \times n$ blocks

$$R = \begin{pmatrix} P & K \\ Q & L \end{pmatrix},$$

then $R^{-1} = -JR^*J$ implies

$$R^{-1} = \begin{pmatrix} L^* & -K^* \\ -Q^* & P^* \end{pmatrix} \tag{2.5}$$

* Received by the editors February 12, 1981, and in revised form May 25, 1981.
† Mathematisches Institut, Universität Würzburg, D 8700 Würzburg, West Germany.

and

$$-Q^*P+P^*Q=0, \qquad -PK^*+KP^*=0, \tag{2.6}$$

$$L^*P-K^*Q=I. \tag{2.7}$$

DEFINITION 2.1. For a complex polynomial $p(z)=\sum a_\nu z^\nu$ we define $\tilde{p}$ by

$$\tilde{p}(z)=\bar{p}(-z)=\sum \overline{a_\nu}(-z)^\nu.$$

Let $\chi(G)$ denote the characteristic polynomial of the matrix G.

We shall need the following result on symplectic transformations which will be discussed in § 6.

THEOREM 2.1. *Let $M\in\mathbb{C}^{2n\times 2n}$ be a Hamiltonian matrix and let the following condition hold*:

(α) *the elementary divisors which belong to pure imaginary eigenvalues of M have even degrees.*

Then the characteristic polynomial of M can be factored into

$$\chi(M)=(-1)^n q\tilde{q}, \tag{2.8}$$

where q and $\tilde{q}$ have only pure imaginary roots in common. There exists a symplectic matrix R such that

$$R^{-1}MR=\begin{pmatrix} T & F \\ 0 & -T^* \end{pmatrix} \tag{2.9}$$

and $\chi(T)=q$ hold.

If M has no eigenvalues with zero real part then one has $(q,\tilde{q})=1$ in (2.8) and $F=0$ in (2.9), which is [2, Thm. 5]. From the matrix R in (2.9) a solution of (1.1) can be obtained (see [2] for the case $F=0$).

THEOREM 2.2. *The following statements are equivalent*:

(i) *The equation* (1.1) *has a solution.*

(ii) *There exists a symplectic matrix*

$$R=\begin{pmatrix} P & K \\ Q & L \end{pmatrix}$$

with P nonsingular which transforms M into

$$R^{-1}MR=\begin{pmatrix} T & F \\ 0 & -T^* \end{pmatrix}.$$

A solution X of (1.1) *yields a matrix*

$$R=\begin{pmatrix} I & 0 \\ X & I \end{pmatrix}$$

for (2.9). *The matrix R in* (2.9) *gives rise to a solution $X=QP^{-1}$ for which*

$$\chi(A+DX)=\chi(T) \tag{2.10}$$

holds.

Proof. If X is a solution, then $\left(\begin{smallmatrix} I & 0 \\ X & I \end{smallmatrix}\right)$ is symplectic and

$$\begin{pmatrix} I & 0 \\ X & I \end{pmatrix}^{-1} M \begin{pmatrix} I & 0 \\ X & I \end{pmatrix}=\begin{pmatrix} A+DX & D \\ 0 & -(A+DX)^* \end{pmatrix}. \tag{2.11}$$

Conversely, if P is nonsingular, then (2.6) implies

$$(P^{-1})^*Q^* = QP^{-1}$$

and $X := QP^{-1}$ is hermitian. Furthermore, (2.7) yields

$$R^{-1}\begin{pmatrix} I & 0 \\ X & I \end{pmatrix} = \begin{pmatrix} P^{-1} & -K^* \\ 0 & P^* \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix} R = \begin{pmatrix} P & K \\ 0 & P^{-*} \end{pmatrix}$$

with $P^{-*} = (P^{-1})^*$. We have

$$\begin{aligned} \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix} M \begin{pmatrix} I & 0 \\ X & I \end{pmatrix} &= \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix} R \begin{pmatrix} T & F \\ 0 & -T^* \end{pmatrix} R^{-1} \begin{pmatrix} I & 0 \\ X & I \end{pmatrix} \\ &= \begin{pmatrix} P & K \\ 0 & P^{-*} \end{pmatrix} \begin{pmatrix} T & F \\ 0 & -T^* \end{pmatrix} \begin{pmatrix} P^{-1} & -K^* \\ 0 & P^* \end{pmatrix} \\ &= \begin{pmatrix} PTP^{-1} & \Delta \\ 0 & -P^{-*}T^*P^* \end{pmatrix} \end{aligned}$$

and therefore in particular $A + DX = PTP^{-1}$ and (2.2). □

Note that Theorems 2.1 and 2.2 do not require $D \geqq 0$.

3. Notation, definitions, lemmas. This section contains prerequisites for the formulation and for the proof of the main result.

DEFINITION 3.1. Let $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{C}^{n\times m}$ be given. The span of the columns of A^iB, $i = 0, 1, \cdots, n-1$, is called the (A, B)-*controllable subspace* of $\mathbb{C}^n$ and is denoted by $C(A, B)$. The pair (A, B) is called *controllable* if $C(A, B) = \mathbb{C}^n$ or equivalently if rank $(B, AB, \cdots, A^{n-1}B) = n$.

The following criterion is due to Hautus.

LEMMA 3.1 [4]. (i) *The pair (A, B) is controllable if and only if*

$$\text{rank } (A - \lambda I, B) = n \tag{3.1}$$

for all eigenvalues λ of A, or equivalently,

(ii) *(A, B) is not controllable if and only if there exists a $y \in \mathbb{C}^n$, such that y^T is a left eigenvector of A and $y^TB = 0$.*

The condition (3.1) gives rise to the following concept [5].

DEFINITION 3.2. An eigenvalue α of A is called B-*controllable* if

$$\text{rank } (A - \alpha I, B) = n \tag{3.2}$$

holds.

Since $C(A, B)$ is an A-invariant subspace, a suitable choice of a basis of $\mathbb{C}^n$ yields the following decomposition theorem.

LEMMA 3.2 [9, p. 99]. *There exists a nonsingular matrix S such that*

$$SAS^{-1} = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}, \qquad SB = \begin{pmatrix} 0 \\ B_2 \end{pmatrix} \tag{3.3}$$

and (A_{22}, B_2) is controllable. If $m = n$ and $B = B^* \geqq 0$, then

$$SBS^* = \begin{pmatrix} 0 & 0 \\ 0 & B_{22} \end{pmatrix}, \qquad B_{22} \geqq 0$$

and (A_{22}, B_{22}) is controllable.

Controllability is preserved under the following operations.

LEMMA 3.3 [13]. *If (A, B) is controllable, then for any $F \in \mathbb{C}^{m\times n}$, for any nonsingular $S \in \mathbb{C}^{n\times n}$ and $T \in \mathbb{C}^{m\times m}$, the pairs $(A+BF, B)$, (SAS^{-1}, SB) and (A, BT) are controllable.*

DEFINITION 3.3. For $A \in \mathbb{C}^{n\times n}$ let $E_i(A)$ denote the subspace of $\mathbb{C}^n$ that is spanned by the generalized eigenvectors belonging to pure imaginary eigenvalues of A,

$$E_i(A) = \bigcup_{\mu\in\mathbb{R}} \ker (A - i\mu I)^n.$$

LEMMA 3.4. *Each pure imaginary eigenvalue of A is B-controllable if and only if*

$$E_i(A) \subset C(A, B). \tag{3.4}$$

Proof. Let α be a pure imaginary eigenvalue of A and let A be transformed into

$$TAT^{-1} = \begin{pmatrix} A_\alpha & 0 \\ 0 & A_r \end{pmatrix}, \qquad A_\alpha \in \mathbb{C}^{p\times p}, \tag{3.5}$$

such that α is the only eigenvalue of A_α and $\det (A_r - \alpha I) \neq 0$. Obviously (3.2) and (3.4) remain valid if A and B are replaced by TAT^{-1} and TB. Therefore we can assume A to be in block diagonal form (3.5). The generalized eigenspace of A corresponding to α is

$$V_p := \{v | v \in \mathbb{C}^n,\ v = (v_1, \cdots, v_p, 0, \cdots, 0)^T\}.$$

It is sufficient to show that (3.2) means $V_p \subset C(A, B)$. Let $B = \binom{B_\alpha}{B_r}$ be partitioned according to (3.5). Then (3.2) is equivalent to

$$\operatorname{rank} (A_\alpha - \alpha I, B_\alpha) = p \tag{3.6}$$

or, because of Lemma 3.1, equivalent to

$$\operatorname{rank} (B_\alpha, A_\alpha B_\alpha, \cdots, A_\alpha^{p-1} B_\alpha) = p.$$

Hence (3.2) implies $V_p \subset \operatorname{span} (B, AB, \cdots, A^{p-1}B) \subset C(A, B)$. If (3.6) does not hold, then there exists a $y \in \mathbb{C}^p$, such that $y^T(A_\alpha - \alpha I) = 0$ and $y^T B = 0$. Thus $y^T A_\alpha^k B = 0$ for all k and $\dim C(A_\alpha, B_\alpha) < p$. Hence there is a $w \in \mathbb{C}^p$ which is not in $C(A_\alpha, B_\alpha)$ and $\binom{w}{0} \in V_p$ is not in $C(A, B)$. □

LEMMA 3.5. *The matrix A_{11} in (3.3) has no pure imaginary eigenvalues, if and only if all pure imaginary eigenvalues of A are B-controllable.*

Proof. We can assume that A and B are in the form (3.3). Let A_{11} be of size q and A_{22} of size r, $q + r = n$. Since (A_{22}, B_2) is controllable, we have for all $\alpha \in \mathbb{C}$

$$\operatorname{rank} (A - \alpha I, B) = \operatorname{rank} (A_{11} - \alpha I) + r$$

and (3.2) holds for all $\alpha \in i\mathbb{R}$ if and only if $\operatorname{rank} (A_{11} - \alpha I) = q$; i.e., A_{11} has no eigenvalues in $i\mathbb{R}$. □

For a uniqueness proof the following result will be used.

LEMMA 3.6 [3, p. 208]. *If $A \in \mathbb{C}^{n\times n}$ and $F \in \mathbb{C}^{n\times n}$ have no eigenvalues in common, then $AX - XF = 0$ implies $X = 0$.*

The following statements are easy to verify.

LEMMA 3.7. *Let $S \in \mathbb{C}^{n\times n}$ be nonsingular and define*

$$\hat{X} := (S^*)^{-1} X S^{-1}, \quad \hat{A} := SAS^{-1}, \quad \hat{D} := SDS^*, \quad \hat{C} := (S^*)^{-1} C S^{-1},$$

$$\hat{M} = \begin{pmatrix} \hat{A} & \hat{D} \\ \hat{C} & -\hat{A}^* \end{pmatrix}.$$

Then X is a solution of (1.1) *if and only if $\hat{X}$ is a solution of*

$$\hat{X}\hat{D}\hat{X}+\hat{X}\hat{A}+\hat{A}^*\hat{X}-\hat{C}=0.$$

The matrix

$$Z:=\begin{pmatrix} S^{-1} & 0 \\ 0 & S^* \end{pmatrix}$$

is symplectic and $\hat{M}=Z^{-1}MZ$. Furthermore $E_i(\hat{A})=SE_i(A)$, $C(\hat{A},\hat{D})=SC(A,D)$ and $\hat{A}+\hat{D}\hat{X}=S(A+DX)S^{-1}$.

LEMMA 3.8 [2]. *Let X_1 and X_2 be two solutions of* (1.1) *and put $U:=X_2-X_1$ and $G_i:=A+DX_i$, $i=1,2$, then*

$$UG_1+G_2^*U=0. \tag{3.7}$$

The nullspace N of U is invariant under G_i, $i=1,2$, and $G_1|_N=G_2|_N$.

4. The condition of Lancaster and Rodman. One of the standard assumptions on (1.1), namely that no eigenvalue of the Hamiltonian matrix M should have a zero real part, has been relaxed by Lancaster and Rodman [7] by introducing the condition (α) on the elementary divisors of M. If (A, D) is controllable then (α) is also necessary for the existence of a solution [7, p. 228]. It will be easy to show that this is still true if only the pure imaginary eigenvalues of A are D-controllable. As in [7] we will use the following observation.

LEMMA 4.1 [7]. *Let G and D be complex $n\times n$ matrices and let D be hermitian, $D\geqq 0$ and (G, D) be controllable. Then all elementary divisors corresponding to pure imaginary eigenvalues of $\left(\begin{smallmatrix} G & D \\ 0 & -G^* \end{smallmatrix}\right)$ have even degrees.*

We state in the following lemma a more general result. Its proof will be self-contained and should provide an easier access to Lemma 4.1.

DEFINITION 4.1. Let $H(z)=(h_{ik}(z))$ be an $n\times n$ matrix of complex rational functions. Then $\tilde{H}$ is defined as

$$\tilde{H}(z)=(\overline{h_{ki}}(-z)).$$

LEMMA 4.2. *Let $G(z)$ and $D(z)$ be $n\times n$ matrices of complex rational functions which have no pole in α, $\alpha\in i\mathbb{R}$, and for which the following assumptions hold:*

(i) $\det G(z)\not\equiv 0\in\mathbb{C}(z)$;
(ii) $D=\tilde{D}$;
(iii) $D(\alpha)\geqq 0$;
(iv) $\operatorname{rank}(G(\alpha), D(\alpha))=n$.

Then the degrees of the elementary divisors of

$$M(z)=\begin{pmatrix} G(z) & D(z) \\ 0 & -\tilde{G}(z) \end{pmatrix}$$

which belong to the characteristic root α are even.

Proof. We localize at α. Let U and V be two invertible matrices in $\mathbb{C}^{n\times n}(z)$ such that U, U^{-1}, V, V^{-1} have no pole in α. We write

$$G(z)\underset{\alpha}{\sim}S(z)$$

if

$$U(z)G(z)V(z)=S(z) \tag{4.1}$$

holds. As $G(z)$ has no poles in α, we can choose U and V in (4.1) such that

$$S(z)=\operatorname{diag}(1,\cdots,1,(z-\alpha)^{k_1},\cdots,(z-\alpha)^{k_r}),\qquad 0<k_1\leqq\cdots\leqq k_r.$$

Then

$$\begin{pmatrix}U(z) & 0\\ 0 & \tilde{V}(z)\end{pmatrix}M(z)\begin{pmatrix}V(z) & 0\\ 0 & \tilde{U}(z)\end{pmatrix}=\begin{pmatrix}S(z) & U(z)D(z)\tilde{U}(z)\\ 0 & -\tilde{S}(z)\end{pmatrix}. \tag{4.2}$$

It is not difficult to verify that for S and $UD\tilde{U}$ the conditions (i)–(iv) are satisfied. Since the elementary divisors corresponding to α are the same for $M(z)$ and the right-hand side of (4.2), we can assume without loss of generality that $G(z)$ is given by

$$G(z)=\begin{pmatrix}I & 0\\ 0 & G_2(z)\end{pmatrix},\qquad G_2(z)=\operatorname{diag}((z-\alpha)^{k_1},\cdots,(z-\alpha)^{k_r}).$$

Then

$$M(z)=\begin{pmatrix}I & 0 & D_1(z) & D_{12}(z)\\ 0 & G_2(z) & D_{21}(z) & D_2(z)\\ 0 & 0 & -I & 0\\ 0 & 0 & 0 & -\tilde{G}_2(z)\end{pmatrix}\underset{\alpha}{\sim}\begin{pmatrix}I & 0\\ 0 & M_2(z)\end{pmatrix}$$

with

$$M_2(z)=\begin{pmatrix}G_2(z) & D_2(z)\\ 0 & -\tilde{G}_2(z)\end{pmatrix}.$$

From (iv) and $G_2(\alpha)=0$ it follows that

$$\operatorname{rank}(D_{21}(\alpha),D_2(\alpha))=r. \tag{4.3}$$

Suppose $D_2(\alpha)$ is singular and $D_2(\alpha)b=0$ and $b\neq 0$. Then

$$(0\quad b^*)D(\alpha)\begin{pmatrix}0\\ b\end{pmatrix}=b^*D_2(\alpha)b=0$$

and (iii) imply $(0\quad b^*)D(\alpha)=0$ and

$$b^*(D_{21}(\alpha),D_2(\alpha))=0,$$

which contradicts (4.3). Therefore $D_2(\alpha)>0$ (positive definite). Because of

$$M_2(z)\underset{\alpha}{\sim}\begin{pmatrix}0 & I\\ \tilde{G}_2(z)D_2^{-1}(z)G_2(z) & 0\end{pmatrix},$$

we can focus on

$$W(z):=\tilde{G}_2(z)D_2^{-1}(z)G_2(z).$$

It follows from $D_2(\alpha)^{-1}>0$ that the principal minors of W are nonzero. Since G_2 is diagonal, the greatest common divisor of all $m\times m$ minors of W is $(z-\alpha)^{e_m}$, where $e_m=\sum_{\rho=1}^{m}2k_\rho$ and $W\underset{\alpha}{\sim}\tilde{G}_2G_2$. Therefore

$$M(z)\underset{\alpha}{\sim}\operatorname{diag}(1,\cdots,1,(z-\alpha)^{2k_1},\cdots,(z-\alpha)^{2k_r}).$$

Since α-equivalence preserves the elementary divisors belonging to the eigenvalue α, this completes the proof. □

The assumption of a constant and semidefinite D in Lemma 4.1 means that the conditions (ii) and (iii) are satisfied. For $G(z)=G-zI$ controllability of (G, D) implies (iv).

5. The main result.

THEOREM 5.1. *Let A, C and D be complex $n\times n$ matrices, C and D hermitian, $D\geqq 0$, and let h be the polynomial determined by*

$$\chi(A)=h\cdot\chi(A|_{C(A,D)}).$$

Let q be a monic polynomial of degree n with at most pure imaginary zeros in common with $\tilde{q}$. Then the following two sets of conditions are equivalent:

(i) *All pure imaginary eigenvalues of A are D-controllable; i.e.,*

$$E_i(A)\subset C(A, D) \tag{5.1}$$

(or h has no pure imaginary eigenvalues) and there exists a hermitian solution X of

$$XDX+XA+A^*X-C=0 \tag{1.1}$$

with

$$\chi(A+DX)=q. \tag{5.2}$$

(ii) *All elementary divisors corresponding to pure imaginary eigenvalues of*

$$M=\begin{pmatrix} A & D\\ C & -A^*\end{pmatrix}$$

have even degree,

$$\chi(M)=(-1)^n q\tilde{q} \tag{2.8}$$

and

$$(h, \tilde{q})=1. \tag{5.3}$$

Moreover, the solution X in (i) *is necessarily uniquely determined.*

Proof. We divide the proof into three parts: (ii) $\Rightarrow$ (i), (i) $\Rightarrow$ (ii) and uniqueness. For the first two parts we will assume, according to Lemmas 3.7 and 3.2

$$A=\begin{pmatrix} A_1 & 0\\ A_{21} & A_2\end{pmatrix},\quad D=\begin{pmatrix} 0 & 0\\ 0 & D_2\end{pmatrix},\quad D_2\geqq 0, \tag{5.4}$$

where the pair (A_2, D_2) is controllable. Then

$$\chi(A|_{C(A,D)})=\chi(A_2)\quad\text{and}\quad h=\chi(A_1).$$

(ii) $\Rightarrow$ (i). Because of Theorem 2.1 there exists a symplectic R, $R=\left(\begin{smallmatrix} P & K\\ Q & L\end{smallmatrix}\right)$, such that (2.9) and $\chi(T)=q$ hold. If no solution X of (1.1) satisfies $\chi(A+DX)=\chi(T)=q$, then P is singular. We will show first that a singular P is not compatible with (5.3). Suppose there is a $y\in\mathbb{C}^n$, $y\neq 0$, such that $Py=0$. Then

$$R\begin{pmatrix} y\\ 0\end{pmatrix}=\begin{pmatrix} 0\\ Qy\end{pmatrix}\quad\text{and}\quad (0\quad y^*)R^{-1}=(-y^*Q^*\quad 0)$$

follows from (2.5). We have as in [2]

$$0=(0\quad y^*)R^{-1}MR\begin{pmatrix} y\\ 0\end{pmatrix}=-y^*Q^*DQy.$$

$D \geqq 0$ implies

$$DQy = 0. \tag{5.5}$$

From

$$MR\begin{pmatrix} y \\ 0 \end{pmatrix} = R\begin{pmatrix} T & F \\ 0 & -T^* \end{pmatrix}\begin{pmatrix} y \\ 0 \end{pmatrix}$$

we obtain

$$\begin{pmatrix} DQy \\ -A^*Qy \end{pmatrix} = \begin{pmatrix} PTy \\ QTy \end{pmatrix}, \tag{5.6}$$

and because of (5.5) we have $PTy = 0$. Hence the nullspace of P is invariant under T and thus contains an eigenvector of T. We can therefore assume

$$Ty = \lambda y \tag{5.7}$$

and (5.6) yields $-A^*Qy = \lambda Qy$. Put $w := Qy$. Then

$$w^*D = 0, \qquad w^*A = -\bar{\lambda} w^*$$

and $w \neq 0$, since R is nonsingular. If w^* is partitioned according to (5.4) into $w^* = (w_1^*, w_2^*)$ then $w_2^* A_2 = -\bar{\lambda} w_2^*$ and $w_2^* D_2 = 0$. Since (A_2, D_2) is controllable, Lemma 3.1 implies $w_2 = 0$. Thus $w_1 \neq 0$ and $w_1^* A_1 = -\bar{\lambda} w_1^*$. Because of (5.7) the number $-\bar{\lambda}$ is also an eigenvalue of $-T^*$. Thus $-\bar{\lambda}$ is a common zero of $\tilde{q}$ and h, which is in contradiction to (5.3).

Let X be a solution with $\chi(A + DX) = q$ and let

$$X = \begin{pmatrix} X_1 & X_{21}^* \\ X_{21} & X_2 \end{pmatrix} \tag{5.8}$$

be partitioned according to (5.4). Then

$$G := A + DX = \begin{pmatrix} A_1 & 0 \\ A_{21} + D_2 X_{21} & A_2 + D_2 X_2 \end{pmatrix} \tag{5.9}$$

and $q = h\chi(A_2 + D_2 X_2)$. Suppose there exists a pure imaginary eigenvalue α of A which is not D-controllable. It follows from Lemma 3.5 that α is a pure imaginary root of $h = \chi(A_1)$. Then α is a zero of q and also of $\tilde{q}$, contrary to $(h, \tilde{q}) = 1$.

(i) $\Rightarrow$ (ii). We assume now that there is a solution X of (1.1) such that (5.2) holds. Let X and G be written as in (5.8) and (5.9). Then

$$G = \begin{pmatrix} A_1 & 0 \\ G_{21} & G_2 \end{pmatrix}, \qquad G_2 = A_2 + D_2 X_2,$$

and

$$\begin{pmatrix} I & 0 \\ X & I \end{pmatrix}^{-1} M \begin{pmatrix} I & 0 \\ X & I \end{pmatrix} = \begin{pmatrix} A_1 & 0 & 0 & 0 \\ G_{21} & G_2 & 0 & D_2 \\ 0 & 0 & -A_1^* & -G_{21}^* \\ 0 & 0 & 0 & -G_2^* \end{pmatrix}.$$

Only the pure imaginary eigenvalues of M are of interest here. Thus only the matrix

$$\begin{pmatrix} G_2 & D_2 \\ 0 & -G_2^* \end{pmatrix}$$

matters. As (A_2, D_2) and hence by Lemma 3.3 also (G_2, D_2) are controllable, the first statement of (ii) follows from Lemma 4.1. Obviously (2.11) implies (2.8). Using Lemma 3.5 again we see from (5.1) that h has no pure imaginary zeros. Because of $h|q$ and since the common roots of q and $\tilde{q}$ (if any) are pure imaginary, we have $(h, \tilde{q}) = 1$.

We now turn to uniqueness. Let X_i, $i = 1, 2$, be two solutions such that

$$\chi(A + DX_i) = q, \qquad i = 1, 2, \tag{5.10}$$

holds and q satisfies (2.8). As in Lemma 3.8 we put $U := X_2 - X_1$ and $G_i := A + DX_i$, $i = 1, 2$. Suppose X_2 and X_1 are distinct; i.e., $U \neq 0$. Without loss of generality (Lemma 3.7) we can assume that

$$U = \begin{pmatrix} V & 0 \\ 0 & 0 \end{pmatrix}, \qquad \det V \neq 0. \tag{5.11}$$

It follows from Lemma 3.8 that G_i has block triangular form

$$G_i = \begin{pmatrix} H_i & 0 \\ \Gamma_i & B \end{pmatrix}, \qquad i = 1, 2.$$

We show first that no eigenvalue of H_i is pure imaginary. From (3.7) we obtain

$$VH_1 + H_2^* V = 0. \tag{5.12}$$

Let D be partitioned corresponding to (5.11) as

$$D = \begin{pmatrix} D_{11} & D_{12} \\ D_{12}^* & D_{22} \end{pmatrix}.$$

Then $G_2 = G_1 + DU$ implies

$$H_2 = H_1 + D_{11} V. \tag{5.13}$$

Suppose there is a pure imaginary eigenvalue $i\mu$ of H_1 and b is a corresponding eigenvector, $H_1 b = i\mu b$, $b \neq 0$. From (5.13) and (5.12) it follows that

$$b^* V D_{11} V b = b^*(VH_2 - VH_1)b = -b^* H_1^* V b - b^* V H_1 b = 0$$

and therefore that

$$D_{11} V b = 0. \tag{5.14}$$

Put $g := Vb$. Then $g \neq 0$, $D_{11} g = 0$ and g^* is a left eigenvector of H_2,

$$g^* H_2 = i\mu g^*.$$

Let the n-vector f be given by $f := \binom{g}{0}$. Then $f^* DF = 0$ and $D \geqq 0$ imply $f^* D = 0$. Furthermore,

$$f^* G_2 = (g^* 0)\begin{pmatrix} H_2 & 0 \\ \Gamma_2 & B \end{pmatrix} = (g^* H_2 \quad 0) = i\mu f^*.$$

On the other hand, $f^* D = 0$ yields

$$f^* G_2 = f^*(A + DX_2) = f^* A.$$

Therefore

$$\operatorname{rank}(A - i\mu I, D) < n$$

and the eigenvalue $i\mu$ of A is not D-controllable, which is a contradiction to (5.1). Hence H_1 and similarly H_2 have no eigenvalues with zero real part.

To complete the proof it has to be shown that $U \neq 0$ is impossible. The assumption (5.10) yields $\chi(G_1) = q$ and $\chi(-G_2^*) = (-1)^n \tilde{q}$. Clearly, $\chi(H_1) | \chi(G_1)$ and $\chi(-H_2^*) | \chi(-G_2^*)$. Recall that q and $\tilde{q}$ have only pure imaginary roots in common and that the polynomials $\chi(H_1)$ and $\chi(-H_2^*)$ have no such roots. Therefore H_1 and $-H_2^*$ have no common eigenvalues. It follows from Lemma 3.6 that the matrix equation (5.12) has only the solution $V = 0$, in contradiction to (5.11). Hence $U = 0$ or $X_1 = X_2$. □

6. Appendix: Symplectic transformations of Hamiltonian matrices. The proof of Theorem 2.1 that will be given in this section relies on normal forms of Hamiltonian matrices. Since those normal forms (see, e.g., [1] and [8]) are perhaps less well known and usually restricted to real matrices, we shall derive the facts we need directly from the theory of regular pencils of hermitian matrices. We recall the following result (see, e.g., [11] for references).

LEMMA 6.1. *Let G and H be hermitian $n \times n$ matrices with G nonsingular. Then there exists a nonsingular $S \in \mathbb{C}^{n \times n}$ such that $S^*(Gz + H)S$ is the direct sum of blocks of the following types* I *and* II:

I. $$\varepsilon D_r(a) = \varepsilon \begin{pmatrix} & & & z+a \\ & & z+a & 1 \\ & \iddots & \iddots & \\ z+a & 1 & & 0 \end{pmatrix},$$

an $r \times r$ matrix with $a \in \mathbb{R}$ and $\varepsilon = \pm 1$;

II. $$\begin{pmatrix} 0 & D_s(b) \\ D_s(\bar{b}) & 0 \end{pmatrix},$$

a $2s \times 2s$ matrix with $b \notin \mathbb{R}$. D_s is defined by I.

A block of type I corresponds to an elementary divisor $(z+a)^r$, $a \in \mathbb{R}$, of $Gz + H$. To each conjugate pair $(z+b)^s$, $(z+\bar{b})^s$ of nonreal elementary divisors of $Gz + H$ there is associated a block of type II.

We shall use the following notation. The $m \times m$ matrices E_m, N_m, $K_m(\lambda)$ are given by

$$E_m = \begin{pmatrix} & & & 1 \\ & & \iddots & \\ & 1 & & \\ 1 & & & \end{pmatrix} = (\delta_{m+1-i,i}), \qquad N_m = \begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix} = (\delta_{i,i+1}),$$

$$K_m(\lambda) = \begin{cases} 0 & \text{for } \lambda \notin \mathbb{R} \\ \operatorname{diag}(1, 0, \cdots, 0) & \text{for } \lambda \in \mathbb{R}. \end{cases}$$

I_m is the $m \times m$ identity matrix and

$$J_{2m} = \begin{pmatrix} 0 & I_m \\ -I_m & 0 \end{pmatrix}.$$

Whenever possible the subscript m will be dropped. If r is even, $r = 2m$, then $D_r(a)$ can be written as

$$D_r(a) = \begin{pmatrix} 0 & D_m(a) \\ D_m(a) & K_m(a) \end{pmatrix}$$

and we can unify types I and II by defining

$$C_{2m}(\lambda)=\begin{pmatrix}0 & D_m(\lambda)\\ D_m(\bar\lambda) & K_m(\lambda)\end{pmatrix}. \tag{6.1}$$

We have $C_{2m}(\lambda)=zE_{2m}+Y_{2m}(\lambda)$ with

$$Y_{2m}(\lambda)=\begin{pmatrix}0 & E_m(\lambda I_m+N_m)\\ E_m(\bar\lambda I_m+N_m) & K_m(\lambda)\end{pmatrix}.$$

Proof of Theorem 2.1. From (2.3) it follows that $M=J^{-1}(-M^*)J$. Therefore eigenvalues of M with nonzero real part λ and $-\bar\lambda$ and elementary divisors $(z-\lambda)^k$ and $(z+\bar\lambda)^k$ appear in pairs. Because of (α) a factorization (2.8) is feasible.

We associate with M the hermitian pencil

$$Jiz+JM. \tag{6.2}$$

Obviously, $(z+\lambda)^k$ is an elementary divisor of (6.2) if and only if $(z-\lambda i)^k$ is an elementary divisor of $zI-M$. Hence (α) implies that the blocks of type I which appear in the normal form of (6.2) are of even size r and can be represented by (6.1). Let S be a matrix which transforms (6.2) into normal form

$$S^*(Jiz+JM)S=\sum_{\tau=1}^{t}{}^{\cdot}\,\varepsilon_\tau C_{2m_\tau}(\lambda_\tau),$$

where $\sum^{\cdot}$ denotes the direct sum. The bounds $\tau=1$ and t will be omitted in the sequel. Then

$$S^*JS=-i\sum{}^{\cdot}\,\varepsilon_\tau E_{2m_\tau} \tag{6.3}$$

and

$$S^*JMS=(S^*JS)(S^{-1}MS)=\sum{}^{\cdot}\,\varepsilon_\tau Y_{2m_\tau}(\lambda_\tau),$$

which yields

$$S^{-1}MS=i\sum{}^{\cdot}\,E_{2m_\tau}Y_{2m_\tau}(\lambda_\tau). \tag{6.4}$$

From $E^{-1}NE=N^T$ follows

$$\begin{pmatrix}I_m & 0\\ 0 & i\varepsilon E_m\end{pmatrix}^*(-i\varepsilon E_{2m})\begin{pmatrix}I_m & 0\\ 0 & i\varepsilon E_m\end{pmatrix}=J_{2m}$$

and

$$\begin{pmatrix}I_m & 0\\ 0 & i\varepsilon E_m\end{pmatrix}^{-1}[iE_{2m}Y_{2m}(\lambda)]\begin{pmatrix}I_m & 0\\ 0 & i\varepsilon E_m\end{pmatrix}=\begin{pmatrix}i\bar\lambda I_m+iN_m & -\varepsilon E_mK_m(\lambda)E_m\\ 0 & -(i\bar\lambda I_m+iN_m)^*\end{pmatrix}$$

$$=:\begin{pmatrix}T_m(\lambda) & F_m(\lambda)\\ 0 & -T_m(\lambda)^*\end{pmatrix},$$

where $F_m(\lambda)$ is hermitian. The $n\times n$ matrix

$$L:=S\sum_{\tau=1}^{t}{}^{\cdot}\begin{pmatrix}I_{m_\tau} & 0\\ 0 & i\varepsilon E_{m_\tau}\end{pmatrix}$$

has the properties

$$L^*JL=\sum{}^{\cdot}\,J_{2m_\tau}$$

and

$$L^{-1}ML = \sum^{\cdot} \begin{pmatrix} T_{m_\tau}(\lambda_\tau) & F_{m_\tau}(\lambda_\tau) \\ 0 & -T_{m_\tau}(\lambda_\tau)^* \end{pmatrix}. \tag{6.5}$$

Let P be the permutation matrix which transforms (6.5) into

$$P^{-1}L^{-1}MLP = \begin{pmatrix} \sum^{\cdot} T_{m_\tau} & \sum^{\cdot} F_{m_\tau} \\ 0 & -\sum^{\cdot} T_{m_\tau}^* \end{pmatrix}.$$

Then

$$P^*L^*JLP = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix} = J.$$

Therefore $R := LP$ is symplectic and

$$R^{-1}MR = \begin{pmatrix} T & F \\ 0 & -T^* \end{pmatrix}.$$

Moreover, if λ_τ is not real then by choosing it suitably from the pair λ_τ, $\bar{\lambda}_\tau$ we can ensure that $\chi(T) = q$. □

Acknowledgment. I should like to thank a referee for valuable remarks.

REFERENCES

[1] A. CIAMPI, *Classification of Hamiltonian linear systems*, Indiana Univ. Math. J., 23 (1973), pp. 513–526.
[2] W. A. COPPEL, *Matrix quadratic equations*, Bull. Austral. Math. Soc., 10 (1974), pp. 377–401.
[3] F. R. GANTMACHER, *Matrizenrechnung*, Bd. 1, 2. Aufl., Deutscher Verlag d. Wissenschaften, Berlin, 1965.
[4] M. L. J. HAUTUS, *Controllability and observability conditions of linear autonomous systems*, Nederl. Akad. Wetensch. Proc. Ser. A, 72 (1969), pp. 443–448.
[5] ———, *Stabilization, controllability and observability of linear autonomous systems*, Nederl. Akad. Wetensch. Proc. Ser. A, 73 (1970), pp. 448–455.
[6] V. KUČERA, *A review of the matrix Riccati equation*, Kybernetika, 9 (1973), pp. 42–61.
[7] P. LANCASTER AND L. RODMAN, *Existence and uniqueness theorems for the algebraic Riccati equation*, Int. J. Control, 32 (1980), pp. 285–309.
[8] A. J. LAUB AND K. MEYER, *Canonical forms for symplectic and Hamiltonian matrices*, Celest. Mech., 9 (1974), pp. 213–238.
[9] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[10] B. P. MOLINARI, *The time-invariant linear-quadratic optimal control problem*, Automatica, 13 (1977), pp. 347–357.
[11] R. C. THOMPSON, *The characteristic polynomial of a principal subpencil of a hermitian matrix pencil*, Lin. Alg. Appl., 14 (1976), pp. 135–177.
[12] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Autom. Control, AC-16 (1971), pp. 621–634.
[13] W. M. WONHAM, *Linear Multivariable Control*, Lecture Notes in Economics and Mathematical Systems 101, Springer-Verlag, Berlin, 1974.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 1, March 1982

0196-5212/82/0301-0002 $01.00/0

SOME RECURRENCE RELATIONS OF RECURSIVE MINIMIZATION*

C. J. K. BATTY†, M. J. PELLING‡ AND D. G. ROGERS§

Abstract. The recursive minimization problem

$$f(n)=\min\left\{\sum_{i=1}^{r} f(a_i)\right\}+g(n),$$

where the minimum is taken over all r-tuples $\mathbf{a}=(a_1,\cdots,a_r)$ of integers a_i such that $0\leqq a_i<n$, $1\leqq i\leqq r$, $\sum_{i=1}^{r} a_i=n$, is studied. Necessary and sufficient conditions on $g(n)$, satisfied by many nonnegative convex sequences, are found for the solution to be given by the recurrence relation

$$f(n)=\sum_{i=1}^{r} f\left(\left[\frac{n+i-1}{r}\right]\right)+g(n).$$

A similar recurrence relation is found for the solution when g satisfies certain concavity conditions.

1. Introduction.

If a sequence $f=\{f(n): n\geqq 0\}$ is defined recursively by

$$f(0)=f(1)=0, \tag{1.1a}$$

$$f(n)=\min_{0<r<n}(f(r)+f(n-r))+n, \qquad n\geqq 2, \tag{1.1b}$$

then it is easy to see inductively firstly that

$$f(2^p+q)=p2^p+q(p+2), \quad p\geqq 1, \quad 0\leqq q\leqq 2^p \tag{1.2}$$

and secondly that the minimum in (1.1b) is attained for $\rho(n)\leqq r\leqq n-\rho(n)$, where

$$\rho(2^p+q)=\max(2^{p-1}, q), \quad p\geqq 1, \quad 0\leqq q\leqq 2^p. \tag{1.3}$$

In particular,

$$f(n)=f\left(\left[\frac{n}{2}\right]\right)+f\left(\left[\frac{n+1}{2}\right]\right)+n \tag{1.4a}$$

$$=f(\rho(n))+f(n-\rho(n))+n, \tag{1.4b}$$

where $[x]$ denotes the integer part of x.

The sequence f of (1.1), (1.2) and (1.4) occurs in a problem of Morris [6] on the sorting of data. The explicit solution (1.2) of (1.1) was given by Carlitz [2] who also considered some generalizations of (1.3). These were in turn special cases of the recurrence relation

$$f(n)=\sum_{i=1}^{r} f\left(\left[\frac{n+i-1}{r}\right]\right)+g(n)=(r-s)f\left(\left[\frac{n}{r}\right]\right)+sf\left(\left[\frac{n}{r}\right]+1\right)+g(n), \tag{1.5}$$

(where $r\geqq 2$ is a fixed integer, and s, depending on n, is given by $n=[n/r]r+s$), which together with the two integer-valued variable version of (1.5),

$$f(n, m)=\sum_{i=1}^{r} f\left(\left[\frac{n+i-1}{r}\right], m\right)+f(n, m-1), \tag{1.6}$$

* Received by the editors February 8, 1980, and in revised form May 27, 1981.
† Department of Mathematics, University of Edinburgh, Edinburgh EH9 3JZ, Scotland.
‡ Balliol College, Oxford, England.
§ 68 Liverpool Road, Watford, Hertfordshire, WD1 8DN, England.

we propose to study here. Equations of this type (again including (1.4a)) have arisen in some work of Pelling and Rogers [7], [8] on the design of electrical circuits. We shall show in § 4 how to obtain explicit solutions of (1.6) for given sequences $g(n) = f(n, 0)$.

As is suggested by (1.1), solutions of (1.5) and (1.6) are closely related to problems of recursive minimization. Given the sequence $\{g(n): n \geqq 0\}$, consider the following recursive definition generalizing (1.1):

$$f(n) = g(n), \qquad n < k, \tag{1.7a}$$

$$f(n) = \min\left\{ \sum_{i=1}^{r} f(a_i) + g(n): 0 \leqq a_i < n, \sum_{i=1}^{r} a_i = n \right\}, \qquad n \geqq k, \tag{1.7b}$$

where $k \geqq 2$ is a fixed integer. If $\phi = \{\phi(n): n \geqq 0\}$ is any sequence such that $\phi(n) \leqq g(n) (n < k)$ and

$$\phi(n) \leqq \sum_{i=1}^{r} \phi(a_i) + g(n) \tag{1.8}$$

whenever $\sum_{i=1}^{r} a_i = n$, then $\phi(n) \leqq f(n)$ for all n. Thus, if the terminology is adapted of Hammersley and Grimmett [4], who initiated a study of (1.7) in the case $k = r = 2$, f is the maximal solution of the generalized r-subadditive inequality (1.8) (see also [1], [7]). The subadditive inequality has been important in the study of physical problems involving cooperative phenomena (see [3], [4], [5] and the references given there).

The question arises as to for what r-tuples $(a_1, \cdots, a_r)$ is the minimum in (1.7b) attained. Hammersley and Grimmett [4] answered this question in a number of special cases for $k = r = 2$. For example, they showed that if $k = r = 2$ and g is convex increasing with $g(0) = g(1) = 0$, then the minimum is attained by taking $a_1 = [n/2]$ and $a_2 = [(n+1)/2]$, while if g is concave increasing with $g(0) = g(1) = 0$, it is attained by taking $a_1 = \rho(n)$, $a_2 = n - \rho(n)$. In this paper we extend these results to more general values of k and r. Thus in § 3 we show that if g is convex and $g([k/r]) = g([k/r]+1) = 0$, then the minimum in (1.7) is attained by taking $a_i = [(n+i-1)/r]$. Indeed, we obtain a precise description of those g for which the minimum is attained by this choice of a_i. In § 5, we exhibit a choice of a_i which attains the minimum whenever g is concave, increasing and nonnegative.

2. The general problem. Throughout this paper, $r \geqq 2$ and $k \geqq 2$ will be fixed integers. Let $\mathscr{S}$ be the set of all sequences $\{g(n): n \geqq 0\}$, and for $n \geqq k$, let $\mathscr{A}_n$ be the set of all r-tuples $\mathbf{a} = (a_1, \cdots, a_r)$ of integers a_i such that

$$0 \leqq a_1 \leqq \cdots \leqq a_r < n, \qquad \sum_{i=1}^{r} a_i = n.$$

An operator T is defined on $\mathscr{S}$ by

$$(Tg)(n) = g(n), \qquad 0 \leqq n < k, \tag{2.1a}$$

$$(Tg)(n) = \min\left\{ \sum_{i=1}^{r} (Tg)(a_i): \mathbf{a} \in \mathscr{A}_n \right\} + g(n), \qquad n \geqq k, \tag{2.1b}$$

so that if f is given by (1.7), then $f = Tg$.

Similarly if $k \geqq r$, we may define $\mathscr{A}'_n$ to be the set of all r-tuples $\mathbf{a}$ in $\mathscr{A}_n$ with $a_1 \geqq 1$, and T' to be the operator given by:

$$(T'g)(n) = g(n), \qquad 0 \leqq n < k$$

$$(T'g)(n) = \min\left\{\sum_{i=1}^{r} (T'g)(a_i) : \mathbf{a} \in \mathscr{A}'_n\right\} + g(n), \quad n \geqq k.$$

For any sequence $\alpha = \{\boldsymbol{\alpha}(n) : n \geqq k\}$ with $\boldsymbol{\alpha}(n) \in \mathscr{A}_n$, we may define an operator P_α on $\mathscr{S}$ by:

$$(P_\alpha g)(n) = g(n), \qquad 0 \leqq n < k,$$

$$(P_\alpha g)(n) = \sum_{i=1}^{r} (P_\alpha g)(\alpha_i(n)) + g(n), \qquad n \geqq k.$$

Note that P_α is linear and bijective, its inverse being given by:

$$(P_\alpha^{-1} f)(n) = f(n), \qquad 0 \leqq n < k, \tag{2.2a}$$

$$(P_\alpha^{-1} f)(n) = f(n) - \sum_{i=1}^{r} f(\alpha_i(n)), \qquad n \geqq k. \tag{2.2b}$$

Although T is bijective, it is only superlinear in the sense that

$$T(\lambda_1 g_1 + \lambda_2 g_2) \geqq \lambda_1 T g_1 + \lambda_2 T g_2$$

for $g_1, g_2 \in \mathscr{S}$ and $\lambda_1, \lambda_2 \geqq 0$.

For f and g in $\mathscr{S}$, put

$$\mathscr{T}_n(g) = \left\{\mathbf{a} \in \mathscr{A}_n : (Tg)(n) = \sum_{i=1}^{r} (Tg)(a_i) + g(n)\right\},$$

$$\mathscr{M}_n(f) = \left\{\mathbf{a} \in \mathscr{A}_n : \sum_{i=1}^{r} f(a_i) \leqq \sum_{i=1}^{r} f(a'_i) \text{ for all } \mathbf{a}' \in \mathscr{A}_n\right\}.$$

Then

$$\mathscr{T}_n(g) = \mathscr{M}_n(Tg). \tag{2.3}$$

Further put

$$\mathscr{S}_\alpha = \{g \in \mathscr{S} : \boldsymbol{\alpha}(n) \in \mathscr{T}_n(g),\ n \geqq k\}, \qquad \mathscr{R}_\alpha = \{f \in \mathscr{S} : \boldsymbol{\alpha}(n) \in \mathscr{M}_n(f),\ n \geqq k\}.$$

By (2.3), $\mathscr{R}_\alpha = T(\mathscr{S}_\alpha)$.

The problem of determining for which $\mathbf{a}$ in $\mathscr{A}_n$ the minimum in (1.7) is attained is now reduced to finding those α such that g belongs to $\mathscr{S}_\alpha$. In order to assist in determining $\mathscr{S}_\alpha$, we note the following lemmas.

LEMMA 2.1. *For any sequence* $\alpha = \{\boldsymbol{\alpha}(n) : n \geqq k\}$ *with* $\boldsymbol{\alpha}(n) \in \mathscr{A}_n$,

$$\mathscr{S}_\alpha = P_\alpha^{-1}(\mathscr{R}_\alpha) = \{g \in \mathscr{S} : P_\alpha g = Tg\}.$$

Proof. Consider g in $\mathscr{S}$, and suppose that $P_\alpha g \in \mathscr{R}_\alpha$ and that $(P_\alpha g)(n) = (Tg)(n)$ for $0 \leqq n < p$, where $p \geqq k$. Then for $\mathbf{a} \in \mathscr{A}_p$

$$\sum_{i=1}^{r} (Tg)(a_i) = \sum_{i=1}^{r} (P_\alpha g)(a_i) \geqq \sum_{i=1}^{r} (P_\alpha g)(\alpha_i(p)) = \sum_{i=1}^{r} (Tg)(\alpha_i(p)).$$

Hence

$$(Tg)(p) = \sum_{i=1}^{r} (Tg)(\alpha_i(p)) + g(p) = \sum_{i=1}^{r} (P_\alpha g)(\alpha_i(p)) + g(p) = (P_\alpha g)(p).$$

It follows by induction that $P_\alpha g = Tg$.

Conversely suppose that $P_\alpha g = Tg$. Then for $\mathbf{a} \in \mathcal{A}_n$, $n \geqq k$,

$$\sum_{i=1}^{r} (P_\alpha g)(a_i) + g(n) = \sum_{i=1}^{r} (Tg)(a_i) + g(n) \geqq (Tg)(n)$$

$$= (P_\alpha g)(n) = \sum_{i=1}^{r} (P_\alpha g)(\alpha_i(n)) + g(n).$$

Thus $\boldsymbol{\alpha}(n) \in \mathcal{M}_n(P_\alpha g)$, so $P_\alpha g \in \mathcal{R}_\alpha$.

The proof that $\{g \in \mathcal{S}: P_\alpha g = Tg\} = T^{-1}(\mathcal{R}_\alpha) = \mathcal{S}_\alpha$ is similar. □

In view of Lemma 2.1 and the simple form of P_α^{-1} given by (2.2), it will suffice in practice to identify $\mathcal{R}_\alpha$. The next two lemmas assist in this process.

LEMMA 2.2. (i). *Let f and f' be sequences in $\mathcal{S}$. For $\mathbf{a} \in \mathcal{M}_n(f)$, $\mathbf{a}' \in \mathcal{M}_n(f')$,*

$$\sum_{i=1}^{r} (f(a_i) - f'(a_i)) \leqq \sum_{i=1}^{r} (f(a_i') - f'(a_i')). \tag{2.4}$$

(ii). *Let f_j, $j \geqq 1$ be sequences in $\mathcal{S}$ such that for any n, $f_j(n) = 0$ for all except finitely many n, and suppose that, for some $p \geqq k$, the sets $\mathcal{M}_p(f_j)$, $j \geqq 1$, have nonempty intersection. Let θ_j, $j \geqq 1$ be strictly positive real numbers, and put $f(n) = \sum_{j \geqq 1} \theta_j f_j(n)$, $n \geqq 0$. Then*

$$\mathcal{M}_p(f) = \bigcap_{j \geqq 1} \mathcal{M}_p(f_j).$$

Proof. (i). Since $\mathbf{a} \in \mathcal{M}_n(f)$,

$$\sum_{i=1}^{r} f(a_i) \leqq \sum_{i=1}^{r} f(a_i').$$

Since $\mathbf{a}' \in \mathcal{M}_n(f')$,

$$\sum_{i=1}^{r} f'(a_i) \geqq \sum_{i=1}^{r} f'(a_i').$$

Now (2.4) follows by subtraction.

(ii). Take fixed $\mathbf{a} \in \bigcap_{j \geqq 0} \mathcal{M}_p(f_j)$. For general $\mathbf{a}' \in \mathcal{A}_p$, $j \geqq 1$,

$$\sum_{i=1}^{r} f_j(a_i') \geqq \sum_{i=1}^{r} f_j(a_i). \tag{2.5}$$

Multiplying (2.5) by θ_j and summing over j gives:

$$\sum_{i=1}^{r} f(a_i') \geqq \sum_{i=1}^{r} f(a_i). \tag{2.6}$$

Thus $\mathbf{a} \in \mathcal{M}_p(f)$. Furthermore,

$$\mathbf{a}' \in \mathcal{M}_p(f) \Leftrightarrow \text{Equality holds in (2.6)}$$

$$\Leftrightarrow \text{Equality holds in (2.5) for all } j$$

$$\Leftrightarrow \mathbf{a}' \in \mathcal{M}_p(f_j) \text{ for all } j. \quad \square$$

Lemma 2.2 shows that $\mathscr{R}_\alpha$ and hence $\mathscr{S}_\alpha = P_\alpha^{-1}(\mathscr{R}_\alpha)$ are convex cones in $\mathscr{S}$.

Define special sequences F_1, F_2, F_3, G_1, G_2 and G_3 by

$$F_1(n) = 1, \qquad F_2(n) = n, \qquad n \geqq 0, \tag{2.7a}$$

$$G_1(n) = 1, \qquad G_2(n) = n, \qquad 0 \leqq n < k, \tag{2.7b}$$

$$G_1(n) = 1 - r, \qquad G_2(n) = 0, \qquad n \geqq k. \tag{2.7c}$$

$$F_3(n) = G_3(n) = \begin{cases} 0, & r > 2, \quad n \geqq 0, \\ 1, & r = 2, \quad n = 0, \\ 0, & r = 2, \quad n > 0. \end{cases} \tag{2.7d}$$

LEMMA 2.3. *Let* λ_1, λ_2 *and* λ_3 *be real numbers,* $f = \lambda_1 F_1 + \lambda_2 F_2 + \lambda_3 F_3$, *and* $g = \lambda_1 G_1 + \lambda_2 G_2 + \lambda_3 G_3$. *Then* $\mathscr{M}_n(f) = \mathscr{A}_n$. *For any sequence* $\alpha = \{\boldsymbol{\alpha}(n): n \geqq k\}$ *of* r*-tuples* $\boldsymbol{\alpha}(n)$ *in* $\mathscr{A}_n$, $P_\alpha g = Tg = f$.

Proof. For $\mathbf{a}$ in $\mathscr{A}_n$, $F_3(a_i) = 0$, so $\sum_{i=1}^r f(a_i) = \lambda_1 r + \lambda_2 n$. Since this is independent of $\mathbf{a}$, it follows immediately that $\mathscr{M}_n(f) = \mathscr{A}_n$. Now $f \in \mathscr{R}_\alpha$ and by (2.2), $P_\alpha g = f$, so by Lemma 2.1, $Tg = P_\alpha g$. □

3. The convex case. Firstly we consider the special choice of r-tuples $\boldsymbol{\beta}(n)$ in $\mathscr{A}_n$ given by

$$\beta_i(n) = \left[\frac{n+i-1}{r}\right], \qquad n \geqq k, \quad 1 \leqq i \leqq r.$$

Then, if we write $f(n, m) = (P_\beta^m g)(n)$, $m, n \geqq 0$, $f(n, m)$ is given by (1.6) for $m \geqq 1$, $n \geqq k$.

Let $l = [k/r]$, so that $lr \leqq k < lr + r$. We define special sequences δ_j, $j \geqq 0$; f_j, $j \geqq 1$ in $\mathscr{S}$ as follows:

$$\delta_j(n) = \begin{cases} 1, & n = j, \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

$$f_j = \delta_{j-1}, \qquad 1 \leqq j \leqq l. \tag{3.2a}$$

$$f_j(n) = \begin{cases} n - k + rj - j, & 0 \leqq n \leqq k - rj + j \\ 0, & k - rj + j \leqq n \leqq j \\ n - j, & j \leqq n \end{cases}, \quad j > l. \tag{3.2b}$$

Note that, for $j > l$, $k - rj + j \leqq l < j$. Hence

$$f_j(n) \geqq n - j, \qquad n \geqq 0, \quad j > l, \tag{3.3}$$

and strict inequality holds in (3.3) unless $n \geqq j$. Furthermore,

$$\begin{aligned} &f_j(l) = f_j(l+1) = 0, \qquad j \geqq 1, \\ &f_j(n) = 0, \qquad j > \left[\frac{k}{r-1}\right], \quad n \leqq l, \\ &f_j(n) = 0, \qquad l < n \leqq j. \end{aligned} \tag{3.4}$$

If $r = 2$, then $l \geqq 1$ and $f_1 = F_3$ as given by (2.7d).

Any sequence f in $\mathscr{S}$ has a unique decomposition as

$$f = \lambda_1(f) F_1 + \lambda_2(f) F_2 + \sum_{j \geqq 1} \theta_j(f) f_j, \tag{3.5}$$

where

$$\lambda_1(f) = (l+1)f(l) - lf(l+1), \tag{3.6a}$$

$$\lambda_2(f) = f(l+1) - f(l), \tag{3.6b}$$

$$\theta_j(f) = f(j+1) - 2f(j) + f(j-1), \qquad j > l, \tag{3.7a}$$

$$\theta_j(f) = f(j-1) - \lambda_1(f) - \lambda_2(f)(j-1) - \sum_{i>l} \theta_i(f) f_i(j-1), \qquad 1 \leqq j \leqq l. \tag{3.7b}$$

(Note that by (3.4) the infinite sum in (3.7b) is essentially finite.) By (3.7a)

$$f \text{ is convex for } n \geqq l \Leftrightarrow \theta_j(f) \geqq 0, \qquad j > l. \tag{3.8a}$$

If f is convex for $n \geqq 0$, then for $1 \leqq j \leqq l$, $f(j-1) \geqq \lambda_1(f) + \lambda_2(f)(j-1)$ and $f_i(j-1) \leqq 0$, $i > l$, so (3.7b) shows that

$$f \text{ is convex for } n \geqq 0 \Rightarrow \theta_j(f) \geqq 0, \qquad j \geqq 1. \tag{3.8b}$$

LEMMA 3.1. *For $j \geqq 1$, $f_j \in \mathscr{R}_\beta$. Furthermore, for $1 \leqq j \leqq l$,*

$$\mathscr{M}_n(f_j) = \{\mathbf{a} \in \mathscr{A}_n : a_i \neq j-1\}, \qquad n \geqq k.$$

For $j > l$,

$$\mathscr{M}_k(f_j) = \{\mathbf{a} \in \mathscr{A}_k : a_r \leqq j \text{ or } a_2 \geqq j\},$$

$$\mathscr{M}_n(f_j) = \{\mathbf{a} \in \mathscr{A}_n : a_r \leqq j\}, \qquad k < n \leqq rj,$$

$$\mathscr{M}_n(f_j) = \{\mathbf{a} \in \mathscr{A}_n : a_1 \geqq j\}, \qquad rj \leqq n.$$

Proof. First suppose $j \leqq l$ and $n \geqq k$. Since $\beta_i(n) \geqq l$, $\sum_{i=1}^r f_j(\beta_i(n)) = 0$. Also for $\mathbf{a} \in \mathscr{A}_n$, $f_j(a_i) \geqq 0$, so

$$\sum_{i=1}^r f_j(a_i) \geqq 0. \tag{3.9}$$

Thus $\boldsymbol{\beta}(n) \in \mathscr{M}_n(f_j)$. Furthermore $\mathbf{a} \in \mathscr{M}_n(f_j)$ if and only if equality holds in (3.9), i.e., $f_j(a_i) = 0$ or $a_i \neq j-1$ $(1 \leqq i \leqq r)$.

Now take $j > l$ and $n \geqq rj$. Then $\beta_i(n) \geqq j$, so $f_j(\beta_i(n)) = \beta_i(n) - j$, and $\sum_{i=1}^r f_j(\beta_i(n)) = n - rj$. For $\mathbf{a}$ in $\mathscr{A}_n$, (3.3) shows that

$$\sum_{i=1}^r f_j(a_i) \geqq \sum_{i=1}^r (a_i - j) = n - rj. \tag{3.10}$$

Thus $\boldsymbol{\beta}(n) \in \mathscr{M}_n(f_j)$, and $\mathbf{a} \in \mathscr{M}_n(f_j)$ if and only if equality holds in (3.10), i.e., $f_j(a_i) = a_i - j$, or $a_i \geqq j$, $1 \leqq i \leqq r$ (see (3.3)).

For $k \leqq n \leqq rj$, we have $l \leqq \beta_i(n) \leqq j$, so $\sum_{i=1}^r f_j(\beta_i(n)) = 0$. For $\mathbf{a} \in \mathscr{A}_n$ with $a_1 \geqq k - rj + j$, $f_j(a_i) \geqq 0$, so $\sum_{i=1}^r f_j(a_i) \geqq 0$. Furthermore, equality holds if and only if $a_i \leqq j$. But if $a_r \leqq j$, then $a_i \leqq a_r \leqq j$ and $a_1 = n - \sum_{i=2}^r a_i \geqq k - rj + j$. For any other $\mathbf{a} \in \mathscr{A}_n$, let s be the largest integer such that $a_s \leqq k - rj + j$, so that $1 \leqq s \leqq r$. Then by (3.3)

$$f_j(a_i) = a_i - (k - rj + j), \qquad 1 \leqq i \leqq s,$$

$$f_j(a_i) \geqq a_i - j, \qquad s < i \leqq r.$$

Hence

$$(3.11)\quad \sum_{i=1}^{r} f_j(a_i) \geqq n - s(k - rj + j) - (r - s)j \geqq k - sk + jr(s-1) = (s-1)(rj - k) \geqq 0,$$

since $s \geqq 1$ and $k < rj$. It now follows that $\boldsymbol{\beta}(n) \in \mathcal{M}_n(f_j)$, and that $\mathbf{a} \in \mathcal{M}_n(f_j)$ if and only if $\sum_{i=1}^{r} f_j(a_i) = 0$. If $a_1 \geqq k - rj + j$, we have already seen that this occurs precisely when $a_r \leqq j$. If $a_1 \leqq k - rj + j$, it occurs precisely when equality holds thoughout (3.11). Thus

$$\mathbf{a} \in \mathcal{M}_n(f_j) \Leftrightarrow n = k, \quad s = 1, \quad f(a_i) = a_i - j, \qquad i > s$$

$$\Leftrightarrow n = k, \quad s = 1, \quad a_2 \geqq j.$$

But if $n = k$ and $a_2 \geqq j$, then it is automatic that

$$a_1 = k - \sum_{i=2}^{r} a_i \leqq k - rj + j,$$

so $s = 1$. This completes the proof. □

We can now obtain our description of $\mathcal{R}_\beta$.

THEOREM 3.2. *A sequence f in $\mathcal{S}$ belongs to $\mathcal{R}_\beta$ if and only if f is of the form*

$$f = \lambda_1 F_1 + \lambda_2 F_2 + \sum_{j \geqq 1} \theta_j f_j,$$

where $\theta_j \geqq 0$ ($j \geqq 1$ if $r > 2$, $j \geqq 2$ if $r = 2$). In this case, for $n > k$, $\mathcal{M}_n(f)$ is the set $\mathcal{A}_n(\theta)$ or r-tuples $\mathbf{a}$ in $\mathcal{A}_n$ such that $\theta_j = 0$ whenever any of the following occurs:

(a) $j = a_i + 1 \leqq l$ *for some* i;

(b) $\max(l, a_1) < j \leqq \beta_1(n)$;

(c) $\beta_r(n) \leqq j < a_r$.

Also $\mathcal{M}_k(f)$ is the set $\mathcal{A}_k(\theta)$ of r-tuples $\mathbf{a}$ in $\mathcal{A}_k$ such that $\theta_j = 0$ whenever either of the following occurs:

(a) $j = a_i + 1 \leqq l$ *for some* i;

(b)′ $\max(l, a_2) < j < a_r$.

Proof. It is immediate from Lemmas 2.2, 2.3 and 3.1 that $f \in \mathcal{R}_\beta$ if f is of the given form.

Conversely suppose $f \in \mathcal{R}_\beta$ and consider $j > l$. Put

$$a_1 = j - 1,$$

$$a_i = j, \qquad 1 < i < r,$$

$$a_r = j + 1.$$

Then $\mathbf{a} \in \mathcal{A}_{rj}$ provided that $j + 1 < rj$. Hence

$$f(j-1) + (r-2)f(j) + f(j+1) \geqq rf(j).$$

By (3.7a), $\theta_j = \theta_j(f) \geqq 0$. The exceptional case occurs only if $r = 2$ and $j = 1$.

Now consider j with $1 \leqq j \leqq l$ and choose the integer $p \geqq l$ such that $k - p(r-1) \geqq j > k - (p+1)(r-1)$. Let $s = j - k + (p+1)(r-1)$, so that $0 < s < r$. Put

$$a_1 = j - 1,$$

$$a_i = p, \qquad 2 \leqq i \leqq s,$$

$$a_i = p + 1, \qquad s < i \leqq r.$$

Then $\mathbf{a} \in \mathscr{A}_k$, provided that $p+1<k$. Furthermore by Lemma 3.1, $\mathbf{a} \in \mathscr{M}_k(f_q)$, $q \neq j$, so by Lemmas 2.2 and 2.3,

$$\theta_j = \sum_{i=1}^{r} f'(a_i) \geqq \sum_{i=1}^{r} f'(\beta_i(k)) = 0,$$

where $f' = f - (\lambda_1 F_1 + \lambda_2 F_2 + \sum_{q \neq j} \theta_q f_q) = \theta_j f_j$. The exceptional case $p+1 \geqq k$ occurs only if $r=2$ and $j=1$.

By Lemmas 2.2 and 2.3,

$$\mathscr{M}_n(f) = \cap \mathscr{M}_n(f_j),$$

where the intersection is taken over those $j \geqq 1$ (if $r>2$) or $j \geqq 2$ (if $r=2$) for which $\theta_j > 0$. It is easily seen from Lemma 3.1 that $\mathscr{M}_n(f) = \mathscr{A}_n(\theta)$. □

It is immediate from Theorem 3.2 and (3.8) that any sequence in $\mathscr{R}_\beta$ is convex for $n \geqq l$, and conversely if f in $\mathscr{S}$ is convex for $n \geqq 0$, then $f \in \mathscr{R}_\beta$. It is possible to give a short direct proof of this latter fact as follows:

Let $\mathbf{a}$ be any r-tuple in $\mathscr{M}_n(f)$, and suppose that $a_r - a_1 > 1$. Let $\mathbf{a}'$ be the r-tuple in $\mathscr{A}_n$ such that

$$\{a_i' : 1 \leqq i \leqq r\} = \left\{\left[\frac{a_1+a_r}{2}\right], \left[\frac{a_1+a_r+1}{2}\right]\right\} \cup \{a_i : 1 < a_i < r\}.$$

Then

$$\sum_{i=1}^{r} f(a_i') = f\left(\left[\frac{a_1+a_r}{2}\right]\right) + f\left(\left[\frac{a_1+a_r+1}{2}\right]\right) + \sum_{i=2}^{r-1} f(a_i) \leqq \sum_{i=1}^{r} f(a_i)$$

by convexity, so $\mathbf{a}' \in \mathscr{M}_n(f)$. After repeated replacements of $\mathbf{a}$ by $\mathbf{a}'$, we eventually reach $\boldsymbol{\beta}(n)$. Thus $\boldsymbol{\beta}(n) \in \mathscr{M}_n(f)$.

Now let $g_j = P_\beta^{-1} f_j$, so that by (2.2), (3.1) and (3.2),

$$g_j = \delta_{j-1} = f_j, \qquad 1 \leqq j \leqq l, \tag{3.12a}$$

$$g_j(n) = \begin{cases} n-k+rj-j, & 0 \leqq n \leqq k-rj+j \\ 0, & k-rj+j \leqq n \leqq j \\ n-j, & j \leqq n \leqq rj \\ rj-j, & rj \leqq n \end{cases}, \quad j > l. \tag{3.12b}$$

Any g in $\mathscr{S}$ has a unique decomposition as

$$g = \mu_1(g) G_1 + \mu_2(g) G_2 + \sum_{j \geqq 1} \gamma_j(g) g_j. \tag{3.13}$$

Here

$$\mu_1(g) = \lambda_1(P_\beta g), \quad \mu_2(g) = \lambda_2(P_\beta g), \quad \gamma_j(g) = \theta_j(P_\beta g). \tag{3.14}$$

It is possible, but not particularly instructive, to give explicit expressions for $\gamma_j(g)$ in terms of the values of g. Instead we note that the decompositions (3.13) for f_i are

$$f_i = g_i, \qquad 1 \leqq i \leqq l,$$

$$f_i = \sum_{j=1}^{l} \left[-\sum_{p \geqq 1} g_{r^p i}(j-1) \right] g_j + \sum_{p \geqq 0} g_{r^p i}, \qquad l < i,$$

so

$$\gamma_j(f_i) \geqq 0, \qquad i, j \geqq 1. \tag{3.15}$$

It follows from (3.5) and the linearity of γ_j that

$$\gamma_j(g) = \lambda_1(g)\gamma_j(F_1) + \lambda_2(g)\gamma_j(F_2) + \sum_{i \geqq 1} \theta_i(g)\gamma_j(f_i). \tag{3.16}$$

THEOREM 3.3. *A sequence g in $\mathcal{S}$ belongs to $\mathcal{S}_\beta$ if and only if g is of the form*

$$g = \mu_1 G_1 + \mu_2 G_2 + \sum_{j \geqq 1} \gamma_j g_j,$$

where $\gamma_j \geqq 0$ ($j \geqq 1$ if $r > 2$, $j \geqq 2$ if $r = 2$). In this case, $\mathcal{T}_n(g) = \mathcal{A}_n(\gamma)$ ($n \geqq k$).

Proof. This follows immediately from Theorem 3.2, Lemma 2.1 and (2.3). □

COROLLARY 3.4. *Let g be a sequence in $\mathcal{S}_\beta$ such that $g(l) = g(l+1) = 0$. Then $T^m g \in \mathcal{S}_\beta$ ($m \geqq 0$).*

Proof. It suffices by induction to prove that $(P_\beta g)(l) = (P_\beta g)(l+1) = 0$, and that $P_\beta g \in \mathcal{S}_\beta$. Provided $l+1 < k$, we have immediately $(P_\beta g)(l) = g(l) = 0$ and $(P_\beta g)(l+1) = g(l+1) = 0$. The exceptional case $l+1 = k$ occurs only if $r = k = 2$, and then $(P_\beta g)(2) = 2g(1) + g(2) = 0$. Thus in all cases $(P_\beta g)(l) = (P_\beta g)(l+1) = 0$.

By (3.6), $\lambda_1(P_\beta g) = \lambda_2(P_\beta g) = 0$. By Theorem 3.3, $\gamma_j(g) \geqq 0$, so by (3.14), $\theta_j(P_\beta g) \geqq 0$. By (3.15) and (3.16), $\gamma_j(P_\beta g) \geqq 0$. By Theorem 3.3, $P_\beta g \in \mathcal{S}_\beta$. □

COROLLARY 3.5. *Let g be a convex sequence in $\mathcal{S}$ such that $g(l) = g(l+1) = 0$. Then $T^m g \in \mathcal{S}_\beta$, $m \geqq 0$. If g is strictly convex for $n \geqq p$ for some $p \geqq l$, then $\mathcal{T}_n(T^m g) = \{\boldsymbol{\beta}(n)\}$, $m \geqq 0$, $n \geqq rp + r - 1$.*

Proof. By (3.6), $\lambda_1(g) = \lambda_2(g) = 0$, and by (3.8b), $\theta_j(g) \geqq 0$. By (3.15) and (3.16), $\gamma_j(g) \geqq 0$. Thus $g \in \mathcal{S}_\beta$ by Theorem 3.3, and $T^m g \in \mathcal{S}_\beta$ by Corollary 3.4.

If g is strictly convex for $n \geqq p$, then by (3.7a), $\theta_j(g) > 0$, $j > p$, so $\gamma_j(g) > 0$, $j > p$. It is now routine to verify from the definition of $\mathcal{A}_n(\gamma)$ that $\mathcal{A}_n(\gamma) = \{\boldsymbol{\beta}(n)\}$, $n \geqq rp + r - 1$, so by Theorem 3.3, $\mathcal{T}_n(g) = \{\boldsymbol{\beta}(n)\}$. Furthermore $Tg = \sum_{j \geqq 1} \gamma_j(g) f_j$ is convex for $n \geqq 0$ and strictly convex for $n \geqq p$, so it follows by induction that $\mathcal{T}_n(T^m g) = \{\boldsymbol{\beta}(n)\}$, $m \geqq 0$, $n \geqq rp + r - 1$. □

In the exceptional case $k = r = 2$, the condition that $g(1) = g(2) = 0$ in Corollaries 3.4 and 3.5 can be relaxed. For

$$F_1 = -G_1 + 2G_2 + 2g_1,$$
$$F_2 = -G_1 + 2G_2 + g_1 + g_2,$$

so $\gamma_j(F_1) = \gamma_j(F_2) = 0$, $j \geqq 3$. Hence if g is any sequence in $\mathcal{S}_\beta$ with $\gamma_2(P_\beta g) = g(1) + g(3) \geqq 0$, then $Tg = P_\beta g \in \mathcal{S}_\beta$ by (3.16) and Theorem 3.3. If in addition, $g(1) \geqq 0$ and $3g(1) + g(2) \geqq 0$, then $T^m g \in \mathcal{S}_\beta$ ($m \geqq 0$). Similarly if g is a convex sequence in $\mathcal{S}$ with $\gamma_2(g) = g(3) - g(2) \geqq 0$, then $g \in \mathcal{S}_\beta$.

COROLLARY 3.6. *Suppose $k \geqq r$ so that $\boldsymbol{\beta}(n) \in \mathcal{A}'_n$, $n \geqq k$, and let g be a sequence in $\mathcal{S}$. Then $T'g = P_\beta g$ if and only if g is of the form*

$$g = \mu_1 G_1 + \mu_2 G_2 + \sum_{j \geqq 1} \gamma_j g_j$$

where $\gamma_j \geqq 0$, $j \geqq 2$.

Proof. This may be proved by making minor amendments to the argument leading to Theorem 3.3. □

In Theorem 3.3 we have described exactly the class $\mathscr{S}_\beta$ of sequences g for which the minimum in (2.1) is attained for $a_i = \beta_i(n) = [(n+i-1)/r]$ for all n, so that $f = Tg$ satisfies (1.5). This choice of $\mathbf{a}$ is a natural one, since it has the least possible spread between its components. But it is also important in practice since $\mathscr{S}_\beta$ includes convex sequences g with $g(n) = 0$ for small enough n (Corollary 3.5). It is rather remarkable that this class is invariant under the minimization process T (Corollary 3.4).

For g in $\mathscr{S}_\beta$, we can find $T^m g$ without performing any minimization, since $T^m g = P_\beta^m g$. In the following section, we consider some methods of calculating directly $P_\beta^m g$, avoiding the iterative procedure involved in the definition of P_β.

4. The operator P_β. Since many convex sequences in $\mathscr{S}$ belong to $\mathscr{S}_\beta$ (Corollary 3.5), it is important to obtain explicit information about the operator P_β. A convenient method of doing this is to use generating functions.

THEOREM 4.1. *Suppose g in $\mathscr{S}$ satisfies $g(n) = 0$, $0 \leqq n \leqq [(k-2)/r]+1$, and let*

$$\Psi(x) = \frac{(1-x)^2}{x} \sum_{n \geqq 0} g(n)x^n, \qquad \Phi(x) = \frac{(1-x)^2}{x} \sum_{n \geqq 0} (P_\beta g)(n)x^n.$$

Then

$$\Phi(x) = \sum_{p \geqq 0} \Psi(x^{r^p}). \tag{4.1}$$

Proof. We write

$$\psi(x) = \sum_{n \geqq 0} g(n)x^n, \qquad \phi(x) = \sum_{n \geqq 0} (P_\beta g)(n)x^n.$$

Then we have, at least formally, noting that (1.5) holds for all $n \geqq 0$ when $f = P_\beta g$,

$$\phi(x) = \sum_{i=1}^{r} \sum_{n \geqq 0} (P_\beta g)\left(\left[\frac{n+i-1}{r}\right]\right)x^n + \psi(x) = \phi(x^r) \sum_{i=1}^{r} h_i(x) + \psi(x), \tag{4.2}$$

where

$$x^i h_i(x) = \sum_{j=1}^{r} x^j = \frac{x(1-x^r)}{1-x}, \qquad 1 \leqq i \leqq r,$$

so that

$$\sum_{i=1}^{r} h_i(x) = \frac{x(1-x^r)}{1-x} \sum_{i=1}^{r} x^{-i} = \frac{1}{x^{r-1}}\left(\frac{1-x^r}{1-x}\right)^2.$$

Hence from (4.2)

$$\phi(x) = \frac{1}{x^{r-1}}\left(\frac{1-x^r}{1-x}\right)^2 \phi(x^r) + \psi(x)$$

or

$$\Phi(x) = \Psi(x) + \Phi(x^r), \tag{4.3}$$

from which (4.1) follows. □

COROLLARY 4.2 *Suppose g in $\mathscr{S}$ satisfies $g(n) = 0$, $0 \leqq n \leqq [(k-2)/r]+1$, and let*

$$\Phi_m(x) = \frac{(1-x)^2}{x} \sum_{n \geqq 0} (P_\beta^m g)(n)x^n, \quad m \geqq 0, \qquad \Psi(x) = \Phi_0(x).$$

Then

$$\Phi_m(x) = \sum_{p \geqq 0} \binom{m+p-1}{m-1} \Psi(x^{r^p}), \qquad m \geqq 1. \tag{4.4}$$

Proof. Applying Theorem 4.1 to $P_\beta^{m-1} g$, we have

$$\Phi_m(x) = \sum_{p \geqq 0} \Phi_{m-1}(x^{r^p}), \qquad m \geqq 1$$

so that, on iteration,

$$\Phi_m(x) = \sum_{p \geqq 0} \sum_{\mathbf{a}} \Psi(x^{r^p}), \qquad m \geqq 1 \tag{4.5}$$

where the second summation is taken over all m-tuples $\mathbf{a} = (a_1, \cdots, a_m)$ of integers a_i such that $0 \leqq a_i \leqq p$, $1 \leqq i \leqq m$, and $\sum_{i=1}^m a_i = p$. Since the number of such m-tuples is $\binom{m+p-1}{m-1}$, we obtain (4.4) from (4.5). □

It follows from (4.4) or by induction that for $m \geqq 0$, and $j > [(k-2)/r] + 1$,

$$(P_\beta^m \delta_j)(jr^p + q) = (P_\beta^m \delta_j)(jr^p - q) = (r^p - q)\binom{p+m-1}{m-1}, \qquad 0 \leqq q \leqq r^p, \quad p \geqq 0, \tag{4.6a}$$

$$(P_\beta^m \delta_j)(n) = 0 \quad \text{otherwise.} \tag{4.6b}$$

Furthermore for $j < l = [k/r]$,

$$P_\beta^m \delta_j = \delta_j. \tag{4.6c}$$

An arbitrary g in $\mathcal{S}$ with $g(l) = g(l+1) = 0$ has a decomposition $g = \sum g(j)\delta_j$, where the summation does not include $j = l$ or $j = l+1$, so (4.6) leads to an expression for $P_\beta^m g$ more easily than the decomposition (3.13).

If $g(n) = 0$, $0 \leqq n \leqq [(k-2)/r] + 1$, the values of $(P_\beta^m g)(r^p)$ may be obtained more directly. For writing

$$(P_\beta^m g)(r^p) = r^p h_m(p), \qquad 0 \leqq m, p,$$

we have

$$h_m(p) = h_m(p-1) + h_{m-1}(p), \qquad 1 \leqq m, p.$$

Hence

$$h_m(p) = \sum_{i=1}^{p} b(m, p-i) h_0(i), \qquad 1 \leqq m, p,$$

where

$$b(m, i) = \binom{m+i-1}{m-1} = \binom{m+i-1}{i}, \qquad 0 \leqq i, \quad 1 \leqq m.$$

In particular if $h_0(p) = 1$, $p \geqq 1$, (so that $k \leqq r^2 - r + 1$), then

$$h_m(p) = b(p, m) = \binom{p+m-1}{m}, \qquad 0 \leqq m, \quad 1 \leqq p.$$

More generally if $h_0(p) = b(p, j)$, $p \geqq 1$, for some $j \geqq 0$, then $h_m(p) = b(p, m+j)$, $m \geqq 0$, $p \geqq 1$, so that $(P_\beta^m g)(r^p)$ may always be determined directly in this way whenever $h_0(p)$ is a polynomial in p.

We remark that (4.3) is of the form

$$\Phi(x) = \Psi(x) + Q(\Phi)(x), \tag{4.7}$$

where

$$Q(\Phi)(x) = \sum_{i=1}^{s} q_i(x)\Phi(x^i).$$

This suggests the formal solution

$$\Phi(x) = (I-Q)^{-1}(\Psi)(x) = \sum_{p\geqq 0} Q^p(\Psi)(x).$$

The more general form (4.7) arises if, for example, we replace P_β by P_α, where $\boldsymbol{\alpha}(n) = ([(p_i n + q_i)/r_i])$ for some p_i dividing r_i, $1 \leqq i \leqq r$, or

$$\boldsymbol{\alpha}(n) = \left(\left[\frac{n}{r}\right], \left[\frac{n}{r}\right], \cdots, \left[\frac{n}{r}\right], n-(r-1)\left[\frac{n}{r}\right]\right).$$

By the results of this and the previous section, we can now write down $T^m g$ immediately when g is convex and $g(l) = g(l+1) = 0$. It is given by taking the appropriate linear combinations of the equations (4.6) with P_β replaced by T. Thus our analysis of the convex case is complete.

5. The concave case. We consider now another special choice of r-tuples $\boldsymbol{\sigma}(n)$ in $\mathscr{A}_n$. Let

$$k_0 = 0, \qquad k_p = (k-1)r^{p-1}, \qquad p > 0;$$
$$j_p = k_{p+1} - k_p, \qquad p \geqq 0.$$

Any integer $n \geqq k$ has a unique decomposition as

$$n = sk_p + (r-s)k_{p+1} + b, \tag{5.1}$$

where $1 \leqq s \leqq r$, $p \geqq 0$ and $0 \leqq b < j_p$, and we put

$$\sigma_i(n) = \begin{cases} k_p, & 1 \leqq i < s, \\ k_p + b, & i = s, \\ k_{p+1}, & s < i \leqq r. \end{cases} \tag{5.2}$$

In the case $k = r = 2$, we have $\sigma_1(n) = \rho(n)$, $\sigma_2(n) = n - \rho(n)$, where $\rho(n)$ is given by (1.3). Hence $f = P_\sigma g$ satisfies (1.4b).

It would be attractive to follow the pattern of § 3 and determine all the extremal elements of $\mathscr{R}_\sigma$ and hence those of $\mathscr{S}_\sigma$, but this task seems intricate, and we do not undertake it. Instead we merely find some particular sequences in $\mathscr{R}_\sigma$.

Consider fixed integers j, p and q with $0 \leqq p \leqq q$ and $1 \leqq j \leqq j_p$. Any integer $n \geqq 0$ has a unique decomposition as

$$n = k_q + s(n)j_p + b(n), \tag{5.3}$$

where $j - j_p \leqq b(n) < j$. Define

$$f_{jpq}(n) = (js(n) + b(n)^+)^+ = \left[\frac{j}{j_p}(n - k_q - b(n)) + b(n)^+\right]^+ \tag{5.4}$$

where $x^+ = \max(x, 0)$. Note that

$$f_{jpq}(n) = 0 \Leftrightarrow n \leqq k_q, \tag{5.5}$$

$$f_{jpq}(n) = \frac{j}{j_p}(n - k_q - b(n)) + b(n)^+ \Leftrightarrow s(n) \geqq 0 \tag{5.6}$$

$$\Leftrightarrow n \geqq k_q + j - j_p.$$

LEMMA 5.1. *For $0 \leqq p \leqq q$ and $1 \leqq j \leqq j_p$, $f_{jpq} \in \mathcal{S}_\sigma$. Furthermore for $k \leqq n \leqq rk_q$,*

$$\mathcal{M}_n(f_{jpq}) = \{\mathbf{a} \in \mathcal{A}_n : a_r \leqq k_q\}.$$

For $n \geqq \max(k, rk_q)$ and $1 \leqq j < j_p$,

$$\mathcal{M}_n(f_{jpq}) = \left\{\mathbf{a} \in \mathcal{A}_n : s(a_1) \geqq 0,\ b(a_i) \geqq 0,\ 1 \leqq i \leqq r,\ \sum_{i=1}^{r} b(a_i) \leqq j\right\}$$

$$\cup \left\{\mathbf{a} \in \mathcal{A}_n : s(a_1) \geqq 0,\ b(a_i) \leqq 0,\ 1 \leqq i \leqq r,\ \sum_{i=1}^{r} b(a_i) \geqq j - j_p\right\}.$$

For $n \geqq \max(k, rk_q)$,

$$\mathcal{M}_n(f_{j_p pq}) = \{\mathbf{a} \in \mathcal{A}_n : a_1 \geqq k_q\}.$$

Proof. For $k \leqq n \leqq rk_q$ and $\mathbf{a} \in \mathcal{A}_n$, we have $\sigma_i(n) \leqq k_q$, so by (5.5),

$$\sum_{i=1}^{r} f_{jpq}(a_i) \geqq 0 = \sum_{i=1}^{r} f_{jpq}(\sigma_i(n)).$$

Thus $\boldsymbol{\sigma}(n) \in \mathcal{M}_n(f_{jpq})$, and $\mathbf{a} \in \mathcal{M}_n(f_{jpq})$ if and only if $f_{jpq}(a_i) = 0$, $1 \leqq i \leqq r$, i.e., $a_r \leqq k_q$ by (5.5).

For $n \geqq \max(k, rk_q)$, we have $\sigma_i(n) \geqq k_q$, so by (5.6),

$$f_{jpq}(\sigma_i(n)) = \frac{j}{j_p}(\sigma_i(n) - k_q - b(\sigma_i(n)) + b(\sigma_i(n))^+. \tag{5.7}$$

Furthermore, since j_p divides $k_{p'} - k_q$, $p' \geqq q$, it follows from (5.2) and (5.3) that for some s with $1 \leqq s \leqq r$,

$$b(\sigma_i(n)) = 0, \qquad i \neq s,$$

$$b(\sigma_s(n)) = b(n).$$

Hence (5.7) gives

$$\sum_{i=1}^{r} f_{jpq}(\sigma_i(n)) = \sum_{i=1}^{r} \frac{j}{j_p}(\sigma_i(n) - k_q) - \frac{j}{j_p} b(n) + b(n)^+ = \frac{j}{j_p}(n - rk_q - b(n)) + b(n)^+. \tag{5.8}$$

For $\mathbf{a} \in \mathcal{A}_n$, writing $b(a_i) = b_i$ and $b(n) = b$, we have

$$\sum_{i=1}^{r} f_{jpq}(a_i) \geqq \sum_{i=1}^{r} \frac{j}{j_p}(a_i - k_q - b_i) + b_i^+ = \frac{j}{j_p}\left(n - rk_q - \sum_{i=1}^{r} b_i\right) + \sum_{i=1}^{r} b_i^+. \tag{5.9}$$

Now

$$\sum_{i=1}^{r} b_i = tj_p + b, \tag{5.10}$$

where

$$t = s(n) - \sum_{i=1}^{r} s(a_i) - \frac{(r-1)}{j_p} k_q,$$

which is an integer. Comparing (5.9) with (5.8), in order to show that $\boldsymbol{\sigma}(n) \in \mathcal{M}_n(f_{jpq})$, it is sufficient to establish that

$$\sum_{i=1}^{r} b_i^+ \geqq b^+ + tj. \tag{5.11}$$

There are three cases, depending on the sign of t.

If $t < 0$, then $b^+ + tj < 0$, so strict inequality holds in (5.11).

If $t = 0$, then (5.10) becomes $\sum_{i=1}^{r} b_i = b$, from which it follows immediately that $\sum_{i=1}^{r} b_i^+ \geqq b^+$, i.e., (5.11) holds. Furthermore there is equality in (5.11) if and only if all the b_i's have the same sign.

If $t > 0$, then (5.10) gives

$$\sum_{i=1}^{r} b_i^+ \geqq \sum_{i=1}^{r} b_i = (t-1)(j_p - j) + j_p + b - j + tj \geqq b^+ + tj,$$

since $t \geqq 1$, $j_p \geqq j$ and $j_p + b - j \geqq b^+$. Thus (5.11) is valid, and equality holds there if and only if $b_i \geqq 0$, $1 \leqq i \leqq r$, $j_p + b - j = b^+$ and either $t = 1$ or $j = j_p$.

This shows that $\boldsymbol{\sigma}(n) \in \mathcal{M}_n(f_{jpq})$. Furthermore $\mathbf{a} \in \mathcal{M}_n(f_{jpq})$ if and only if equality holds in (5.9) and in (5.11), i.e., $s(a_1) \geqq 0$ by (5.6), and one of the following holds;

(a) $\sum_{i=1}^{r} b_i = b$ and all the b_i's have the same sign;

(b) $j = j_p$, $b \geqq 0$ and $b_i \geqq 0$, $1 \leqq i \leqq r$;

(c) $t = 1$, $b = j - j_p$ and $b_i \geqq 0$, $1 \leqq i \leqq r$.

Thus $\mathcal{M}_n(f_{jpq})$ is as asserted. □

Now let $g_{jpq} = P_\sigma^{-1} f_{jpq}$ so that, by (2.2), (5.2), (5.4) and (5.5),

$$g_{jpq}(n) = \begin{cases} f_{jpq}(n), & 0 \leqq n < \max(k, rk_q), \\ \dfrac{(r-1)k_q j}{j_p}, & \max(k, rk_q) \leqq n. \end{cases} \tag{5.12}$$

THEOREM 5.2. *Let g be a sequence in $\mathcal{S}$ of the form*

$$g = \mu_1 G_1 + \mu_2 G_2 + \sum_{p \geqq 0} \sum_{q \geqq p} \sum_{j=1}^{j_p} \gamma_{jpq} g_{jpq}, \tag{5.13}$$

where $\gamma_{jpq} \geqq 0$. Then $g \in \mathcal{S}_\sigma$. If in addition, $\mu_1 \geqq 0$ and $\mu_2 \geqq 0$, then $T^m g \in \mathcal{S}_\sigma$, $m \geqq 0$.

Proof. It follows immediately from Lemmas 2.1, 2.2, 2.3 and 5.1 that $g \in \mathcal{S}_\sigma$ and

$$Tg = P_\sigma g = \mu_1 F_1 + \mu_2 F_2 + \sum_{p \geqq 0} \sum_{q \geqq p} \sum_{j=1}^{j_p} \gamma_{jpq} f_{jpq}. \tag{5.14}$$

The explicit expressions (2.7), (5.4) and (5.12) show that

$$F_1 = G_1 + r g_{111}, \tag{5.15a}$$

$$F_2 = f_{k_1 00}, \tag{5.15b}$$

$$f_{jpq} = \sum_{q' \geqq q} g_{jpq'}. \tag{5.15c}$$

Thus if $\mu_1 \geqq 0$ and $\mu_2 \geqq 0$, then from (5.14) and (5.15), Tg has a decomposition of the form (5.13) in which the coefficients are non-negative. Thus it follows inductively that $T^m g \in \mathcal{S}_\sigma$, $m \geqq 0$. □

In the special case when $p = q > 0$, we have

$$g_{jpp}(n) = \begin{cases} 0, & 0 \leqq n \leqq k_p, \\ n - k_p, & k_p \leqq n \leqq k_p + j, \\ j, & k_p + j \leqq n. \end{cases}$$

In particular, g_{jpp} is increasing and concave on each of the intervals $k_q \leqq n \leqq k_{q+1}$, $q \geqq 0$. Also

$$g_{j00}(n) = \begin{cases} n, & 0 \leqq n \leqq j, \\ j, & j \leqq n < k, \\ 0, & k \leqq n. \end{cases}$$

Any sequence g in $\mathcal{S}$ has a unique decomposition

$$g = \mu G_1 + \sum_{p \geqq 0} \sum_{j=1}^{i_p} \gamma_{jp} g_{jpp}, \tag{5.16}$$

where

$$\mu = g(0), \tag{5.17}$$

$$\gamma_{11} = 2g(k) - g(k+1) + (r-1)g(0), \tag{5.18a}$$

$$\gamma_{i_p p} = g(k_{p+1}) - g(k_{p+1} - 1), \qquad p \geqq 0, \tag{5.18b}$$

$$\gamma_{jp} = 2g(k_p + j) - g(k_p + j - 1) - g(k_p + j + 1) \quad \text{otherwise.} \tag{5.18c}$$

COROLLARY 5.3. *Let g be a sequence in $\mathcal{S}$, and suppose that $g(0) \geqq 0$ and g is increasing and concave on each of the intervals $k_p \leqq n \leqq k_{p+1}$, $p \geqq 0$. Then $T^m g \in \mathcal{S}_\sigma$, $m \geqq 0$. If g is strictly increasing and strictly concave on each of the intervals, then $\mathcal{T}_n(T^m g) = \{\boldsymbol{\sigma}(n)\}$, $n \geqq k$, $m \geqq 0$.*

Proof. In the decomposition (5.16), we have $\mu \geqq 0$, $\gamma_{jp} \geqq 0$ by (5.17) and (5.18). By Theorem 5.2, $T^m g \in \mathcal{S}_\sigma$, $m \geqq 0$.

If g is strictly increasing and strictly concave on each of the intervals, then g has a decomposition of the form (5.13) in which $\mu_1 \geqq 0$, $\mu_2 \geqq 0$, $\gamma_{jpq} \geqq 0$ and $\gamma_{jpp} > 0$. By Lemmas 2.1, 2.2 and 5.1,

$$\mathcal{T}_n(g) \subset \bigcap_{j,p} \mathcal{T}_n(g_{jpp}) = \{\boldsymbol{\sigma}(n)\}.$$

Furthermore by (5.14) and (5.15), Tg also has a decomposition of the form (5.13) in which $\mu_1 \geqq 0$, $\mu_2 \geqq 0$, $\gamma_{jpq} \geqq 0$ and $\gamma_{jpp} > 0$. It follows that $\mathcal{T}_n(Tg) = \{\boldsymbol{\sigma}(n)\}$, and by induction that $\mathcal{T}_n(T^m g) = \{\boldsymbol{\sigma}(n)\}$. □

Now consider a more general situation in which $\{k_p : p \geqq 0\}$ is a strictly increasing sequence of nonnegative integers such that $k_1 < k$ and $rk_0 \leqq k$, and let $j_p = k_{p+1} - k_p$ ($p \geqq 0$). Then $\boldsymbol{\sigma}(n)$ in $\mathcal{A}_n$ may again be defined by (5.1) and (5.2), and sequences f_{jpq}, $0 \leqq p \leqq q$, $1 \leqq j \leqq j_p$, by (5.3) and (5.4). Assume that j_p divides both j_q and $(r-1)k_q$, $0 \leqq p \leqq q$. Then the statement and proof of Lemma 5.1 go through verbatim. Furthermore, if $g_{jpq} = P_\sigma^{-1} f_{jpq}$, then g_{jpq} is again given by (5.12). If l is the largest integer such

that $rk_l \leqq k$, then

$$T\delta_j = P_\sigma \delta_j = \delta_j, \qquad 0 \leqq j < k_l$$

(cf. (3.12a)). Thus any positive linear combination of the sequences $\pm G_1$, $\pm G_2$, δ_j, $0 \leqq j < k$ and g_{jpq}, $0 \leqq p \leqq q$, $1 \leqq j \leqq j_p$ belongs to $\mathscr{S}_\sigma$.

Suppose further that there is a function π such that $rk_q = k_{\pi(q)}$. Then for $q > l$,

$$f_{jpq} = \sum_{i \geqq 0} g_{jp\pi^i(q)} \in \mathscr{S}_\sigma.$$

Thus $T^m g_{jpq} \in \mathscr{S}_\sigma$, $m \geqq 0$, $q > l$.

The choice $k_0 = 0$, $k_p = (k-1)r^{p-1}$, $p > 0$, was covered above. If we take $k_p = p$, $p \geqq 0$, all the assumptions are satisfied, since $j_p = 1$ and $rk_q = k_{rq}$. In this case,

$$\sigma_i(n) = \left[\frac{n+i-1}{r}\right] = \beta_i(n)$$

as defined in § 3. The functions f_{jpq}, $j = 1$, $0 \leqq p \leqq q$, are now given by

$$f_{jpq}(n) = \begin{cases} 0, & n \leqq q, \\ n-q, & n \geqq q. \end{cases}$$

Thus for $q > l$, (3.7) gives

$$\theta_q(f_{jpq}) = 1,$$
$$\theta_i(f_{jpq}) = -f_q(i-1) \geqq 0, \qquad 1 \leqq i \leqq l,$$
$$\theta_i(f_{jpq}) = 0 \quad \text{otherwise.}$$

Thus the fact that f_{jpq} lies in $\mathscr{S}_\sigma$ for $q > l$ is in this case a consequence of Theorem 3.2. The additional strength of that theorem lies in the fact that any sequence f in $\mathscr{R}_\sigma$ with $f(n) = 0$, $0 \leqq n \leqq l+1$, is a positive linear combination of f_{jpq}, $q > l$.

Finally suppose that $k \geqq r$ and consider the choice

$$k_0 = 1, \qquad k_p = (k-1)r^{p-1}, \qquad p \geqq 1.$$

Although this does not satisfy the condition that j_0 should divide j_q and $(r-1)k_q$, it is still possible to show that $T'g = P_\sigma g$ for any sequence g with $g(1) \geqq 0$ which is increasing and concave on each of the intervals $k_p \leqq n \leqq k_{p+1}$.

One might not necessarily expect the r-tuples $\boldsymbol{\sigma}(n)$ to be of special interest, but Corollary 5.3 shows why they are. If g is nonnegative, concave and increasing, the minimum in (2.1) is attained for $\mathbf{a} = \boldsymbol{\sigma}(n)$, and the same applies when g is replaced by $T^m g$. Thus again we can calculate $T^m g$ without any minimization process, since $T^m g = P_\sigma^m g$.

REFERENCES

[1] C. J. K. BATTY AND D. G. ROGERS, *Some maximal solutions of the generalized subadditive inequality*, preprint.

[2] L. CARLITZ, *A sorting function*, Duke Math. J., 38 (1971), pp. 561–568.

[3] J. M. HAMMERSLEY, *On the rate of convergence to the connective constant of the hypercubical lattice*, Quart. J. Math. Oxford Ser. (2), 12 (1961), pp. 250–256.

[4] J. M. HAMMERSLEY AND G. R. GRIMMETT, *Maximal solutions of the generalized subadditive inequality*, in Stochastic Geometry, E. F. Harding and D. G. Kendall, eds., John Wiley, New York, 1974.

[5] J. M. HAMMERSLEY AND V. V. MENON, *A lower bound for the monomer dimer problem*, J. Inst. Math. Appl., 6 (1970), pp. 341–364.

[6] R. MORRIS, *Some theorems on sorting*, SIAM J. Appl. Math., 17 (1967), pp. 1–6.

[7] M. J. PELLING AND D. G. ROGERS, *A problem in the design of electrical circuits*, in Combinatorial Mathematics V: Proc. 5th Australian Conf., Lecture Notes in Mathematics 622, Springer-Verlag, Berlin, 1977, pp. 153–169.

[8] ———, *Further results on a problem in the design of electrical circuits*, in Combinatorial Mathematics: Proceedings of International Conference, Canberra, 1977, Lecture Notes in Mathematics 686, Springer-Verlag, Berlin, 1978, pp. 240–247.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 1, March 1982

0196-5212/82/0301-0003 $01.00/0

COUNTING MATRICES BY DRAZIN INDEX*

J. V. BRAWLEY†

Abstract. Let $F_q^{n\times n}$ denote the algebra of $n\times n$ matrices over F_q, the finite field of q elements and for each $A\in F_q^{n\times n}$, let Ind(A) denote the Drazin index of A; i.e., Ind (A) is the least nonnegative integer k such that the system of matrix equations (i) $A^{k+1}X=A^k$, (ii) $XAX=X$ and (iii) $AX=XA$ has a (necessarily unique) solution. This paper determines for each $k\geqq 0$ the number of matrices $A\in F_q^{n\times n}$ with Ind $(A)=k$. These results are then extended to cover a more general class of finite rings including the ring Z/Zm of integers modulo m.

1. Introduction. Let $F^{n\times n}$ denote the algebra of $n\times n$ matrices over the field F and let $A\in F^{n\times n}$. It is well known [5] that there exists a unique smallest integer Ind $(A)\geqq 0$ and a unique matrix X, also depending on A, such that X satisfies the matrix equations

$$A^{k+1}X=A^k,\quad XAX=X,\quad AX=XA \tag{1}$$

for all $k\geqq$ Ind (A). (Here, it is understood that $A^0=I$.) The matrix X, alternatively denoted by A^D, is called the *Drazin inverse* of A and the integer Ind (A) is called the *Drazin index* of A. (Note that A is invertible if and only if Ind $(A)=0$.)

The Drazin inverse for matrices over the real or complex number field has many important applications (see, e.g., [3]). Recently Hartwig [9] and Levine and Hartwig [13] have applied the concept of the Drazin inverse for matrices over finite fields and residue class rings of integers to the Hill cryptographic system. Because of this application to cryptography, Hartwig [10] has asked for the number of matrices which have group inverses, i.e., the number of matrices $A\in F^{n\times n}$ which are members of some multiplicative group (within the multiplicative semigroup $F^{n\times n}$). The set of matrices with group inverses can also be described as the set of matrices with Ind $(A)\leqq 1$ or as the set of matrices in the range of the mapping which takes each $n\times n$ matrix A to its Drazin inverse A^D (see [3]).

In the present paper we determine for each integer $k\geqq 0$, the number of $n\times n$ matrices A over a finite field with Drazin index equal to k. We then generalize these results to a larger class of finite rings which include the residue class rings of integers.

2. The Drazin index and nilpotent matrices. For the remainder of this paper F_q will denote the finite field of q elements. It is clear from (1) that a matrix $A\in F_q^{n\times n}$ has Drazin index Ind $(A)=0$ if and only if A is invertible; i.e., A is a member of the general linear group $GL(n, q)$. Thus, the number of A with Ind $(A)=0$ is the number $\gamma(n, q)$ of elements in $GL(n, q)$ which is well known [4] to be

$$\gamma(n, q)=(q^n-1)(q^n-q)\cdots(q^n-q^{n-1}). \tag{2}$$

We can therefore restrict our attention to singular matrices in which case Ind $(A)\geqq 1$.

Consider such a singular matrix A. From the theory of similarity of matrices over fields (see, e.g. [15]), there exists some unique t, $1\leqq t\leqq n$, such that A is similar to a matrix of the form diag (B, N) where $B\in GL(n-t, q)$ and N is a $t\times t$ nilpotent matrix. The Drazin index of A is known to equal the index of nilpotency of N (see [3]); i.e.,

* Received by the editors July 21, 1980, and in revised form May 26, 1981. This research was supported in part by the National Science Foundation under grant ISP-8011451.

† Department of Mathematical Sciences, Clemson University, Clemson, South Carolina 29631. A portion of this work was completed while the author was a Visiting Professor in the Department of Mathematics, University of Tennessee, Knoxville.

Ind (A) is the least positive integer such that $N^k = 0$. Thus $0 \leqq \text{Ind}\,(A) \leqq n$ for every $A \in F_q^{n\times n}$, and the matrices A with $\text{Ind}\,(A) = k \geqq 1$ are precisely those A similar to a matrix of the form diag (B, N) for some invertible matrix B and nilpotent matrix N whose nilpotency index is k. This fact will be used in the next section.

For purposes of enumeration we shall need a formula for the number $\eta(t, q; k)$ of $t \times t$ nilpotent matrices over F_q with nilpotency index k, $1 \leqq k \leqq t$. Fine and Herstein [7] and Gerstenhaber [8] have determined the total number of $t \times t$ nilpotent matrices over F_q to be q^{t^2-t} and both Bollman and Ramirez [1] and Lusztig [14] have enumerated these by rank. In order to enumerate the nilpotent matrices by index of nilpotency one can use the methods of [7] or more simply apply the results of Hodges [12] who counts the number of matrices $A \in F_q^{n\times n}$ satisfying a given polynomial equation $f(x) = 0$, $f(x) \in F_q[x]$. For completeness and to introduce the notation we now give a brief derivation of this number.

Let

$$\pi = [1^{f_1}, 2^{f_2}, \cdots, k^{f_k}] \tag{3}$$

denote a partition of the integer t into parts $1, 2, \cdots, k$ repeated $f_1, f_2, \cdots, f_k$ times, respectively, where $f_i \geqq 0$, $i = 1, 2, \cdots, k-1$ and $f_k > 0$. Then $t = \sum_{i=1}^{k} i \cdot f_i$, and the largest part in the partition is k. Let $\Pi(t, k)$ denote the set of all such partitions of t and associate with each partition $\pi = [1^{f_2}, 2^{f_2}, \cdots, k^{f_k}]$ the nilpotent matrix $J(\pi) =$ diag $(J_1(f_1), J_2(f_2), \cdots, J_k(f_k))$, where for each i, $1 \leqq i \leqq k$, $J_i(f_i)$ is the block diagonal matrix diag $(J_i, J_i, \cdots, J_i)$ with J_i, the $i \times i$ Jordan matrix with 1's on the superdiagonal and 0's elsewhere, being repeated f_i times. Each $t \times t$ nilpotent matrix of nilpotency index k is similar to one and only one $J(\pi)$ for some unique $\pi \in \Pi(t, k)$. The number $S(\pi)$ of nilpotent matrices similar to a given $J(\pi)$ is $\gamma(t, q)/C(\pi)$, where $C(\pi)$ is the number of invertible $t \times t$ matrices commuting with $J(\pi)$ (see [12]). Hence

$$\eta(t, q; k) = \sum_{\pi} \frac{\gamma(t, q)}{C(\pi)} \tag{4}$$

where the sum is over all $\pi \in \Pi(t, k)$. Now Dickson [4, p. 235] has determined the number $C(\pi)$ to be

$$C(\pi) = q^{e(\pi)} \gamma(f_1, q)\gamma(f_2, q) \cdots \gamma(f_k, q), \tag{5}$$

where $\gamma(n, q)$ is given by (2) and where the exponent $e(\pi)$ is

$$e(\pi) = \sum_{j=1}^{k} \sum_{i=1}^{k} f_i f_j \min\,(i, j) - \sum_{i=1}^{k} f_i^2, \tag{6}$$

with min (i, j) denoting the minimum of i and j. Thus, putting these results together we have

$$\eta(t, q; k) = \sum_{\pi} \frac{\gamma(t, q)}{q^{e(\pi)} \gamma(f_1, q)\gamma(f_2, q) \cdots \gamma(f_k, q)}. \tag{7}$$

Formula (7) is essentially the result of Hodges [12, p. 293] applied to our situation.

We note that (7) is especially simple when $k = 1$ for then $\eta(t, q; 1) = 1$ as 0 is the only nilpotent matrix of nilpotency index 1.

3. Counting matrices over finite fields by Drazin index. Let k be an integer $1 \leqq k \leqq n$. We seek the number of $A \in F_q^{n\times n}$ with Drazin index $\text{Ind}\,(A) = k$. To this end, for each integer m, let $\mathscr{C}_m = \{B_1, B_2 \cdots\}$ denote a set of distinct representatives for the similarity (conjugacy) classes of $GL(m, q)$. Then if $C(B)$ denotes the number

of $m \times m$ invertible matrices commuting with B it follows that

$$\gamma(m, q) = \sum_{B \in \mathscr{C}_m} \gamma(m, q)/C(B), \tag{8}$$

as $\gamma(m, q)/C(B)$ is just the number of elements in the similarity class of B. (Formula (8) is the well-known *class equation* for groups applied to $GL(m, q)$.)

Also, a matrix $A \in F_q^{n \times n}$ has Drazin index k if and only if there exist a unique integer t, $k \leqq t \leqq n$, a unique partition $\pi \in \Pi(t, k)$, and a unique matrix $B \in \mathscr{C}_{n-t}$ such that A is similar to

$$D = \operatorname{diag}(B, J(\pi)). \tag{9}$$

Moreover, it is readily verified that a matrix $P \in GL(n, q)$ commutes with the matrix $D = \operatorname{diag}(B, J(\pi))$ of (9) if and only if $P = \operatorname{diag}(P_1, P_2)$ where $P_1 \in GL(n-t, q)$ and commutes with B, and where $P_2 \in GL(t, q)$ and commutes with $J(\pi)$. Thus, the number of matrices of Drazin index k which are similar to $\operatorname{diag}(B, J(\pi))$ is $\gamma(n, q)/C(B)C(\pi)$ where $C(B)$ and $C(\pi)$ denote respectively the number of invertible matrices commuting with B and $J(\pi)$ (see [12, p. 292]). It follows that the number $\delta(n, q; k)$ of matrices in $F_q^{n \times n}$ with Drazin index $k (1 \leqq k \leqq n)$ is

$$\delta(n, q; k) = \sum_{t=k}^{n} \sum_{\pi} \sum_{B} \frac{\gamma(n, q)}{C(B)C(\pi)},$$

where π ranges over $\Pi(t, k)$ and B ranges over $\mathscr{C}_{n-t}$. But using (8) and also (4) we have for fixed t

$$\begin{aligned}
\sum_{\pi} \sum_{B} \frac{\gamma(n, q)}{C(B)C(\pi)} &= \sum_{\pi} \frac{\gamma(n, q)}{C(\pi)\gamma(n-t, q)} \sum_{B} \frac{\gamma(n-t, q)}{C(B)} \\
&= \sum_{\pi} \frac{\gamma(n, q)}{C(\pi)} \\
&= \frac{\gamma(n, q)}{\gamma(t, q)} \sum_{\pi} \frac{\gamma(t, q)}{C(\pi)} \\
&= \frac{\gamma(n, q)\eta(t, q; k)}{\gamma(t, q)}.
\end{aligned}$$

Hence

$$\delta(n, q; k) = \gamma(n, q) \sum_{t=k}^{n} \frac{\eta(t, q; k)}{\gamma(t, q)}, \tag{10}$$

where $\eta(t, q; k)$ is given in (7) and where $\gamma(n, q)$ and $\gamma(t, q)$ are given in (2).

For $k = 1$, since $\eta(t, q; k) = 1$, equation (10) reduces to

$$\delta(n, q; 1) = \sum_{t=1}^{n} \frac{\gamma(n, q)}{\gamma(t, q)}; \tag{11}$$

hence, the number of matrices which have group inverses is

$$\delta(n, q; 0) + \delta(n, q; 1) = \sum_{t=0}^{n} \frac{\gamma(n, q)}{\gamma(t, q)}, \tag{12}$$

where $\gamma(0, q) = 1$ by definition. For finite fields, formula (12) answers the question of Hartwig mentioned in § 1.

4. Generalizations to Fitting rings. For any finite ring R with identity we shall use $\gamma(n, R)$ to denote the number of invertible matrices in $R^{n\times n}$ and $\eta(n, R; k)$ to denote the number of nilpotent matrices in $R^{n\times n}$ whose nilpotency index is k. The number $\gamma(n, R)$ has been determined by Farahat [6] and independently by Brawley [2]. The numbers $\eta(n, R; k)$ are apparently not known for a general ring R.

DEFINITION. A ring R with identity is called a *Fitting* ring if every matrix A over R is similar (over R) to a matrix of the form diag (B, N) where B is invertible and N is nilpotent.

Examples of Fitting rings are fields and the ring Z/Zp^m of integers modulo a prime power p^m [11]. A direct sum of two or more Fitting rings is not a Fitting ring.

Using techniques similar to those in the previous section one can prove the following theorem (whose proof will be omitted).

THEOREM. *Let R be a finite Fitting ring with identity. Then the number $\delta(n, R; k)$ of matrices $A \in R^{n\times n}$ with Drazin index equal to k is $\gamma(n, R)$ if $k = 0$ and*

$$\delta(n, R; k) = \gamma(n, R) \sum_{t=1}^{n} \frac{\eta(t, R; k)}{\gamma(t, R)} \quad \text{if } k \geqq 1. \tag{13}$$

We should comment that every matrix A over a finite ring is a strongly π-regular element and hence A has a Drazin inverse and a Drazin index [5, p. 510].

Since $\eta(t, R; 1) = 1$, we immediately have the following corollary.

COROLLARY. *Let R be a finite Fitting ring with identity. The number of matrices in the range of the mapping on $R^{n\times n}$ which takes each matrix to its Drazin inverse is*

$$\delta(n, R; 0) + \delta(n, R; 1) = \sum_{t=0}^{n} \frac{\gamma(n, R)}{\gamma(t, R)}, \tag{14}$$

where $\gamma(0, R) = 1$ by definition.

We comment that the reason the variable of summation t in (13) ranges from 1 to n rather than k to n as in (10) is because a $t \times t$ nilpotent matrix over R may have an index of nilpotency greater than t.

Finally consider a finite ring R which is the direct sum of Fitting rings; say $R = R_1 \oplus R_2 \oplus \cdots \oplus R_r$. (An example of such a ring is the ring Z/Zm of integers modulo an arbitrary integer m.) If $A \in R^{n\times n}$, then $A = A_1 \oplus \cdots \oplus A_r$ where $A_i \in R_i^{n\times n}$. Now the Drazin index of A is $\max\{k_1, \cdots, k_r\}$ where k_i is the Drazin index of A_i. Thus A has Drazin index k if and only if Ind $(A_i) \leqq k$ for all i with equality holding in at least one component. Hence

$$\sigma(n, R; k) = \sum_{i_1=0}^{k} \cdots \sum_{i_r=0}^{k} \delta(n, R_1; i_1) \cdots \delta(n, R_r; i_r) \tag{15}$$

is the number of $A \in R^{n\times n}$ with Drazin index less than or equal k and

$$\delta(n, R; k) = \sigma(n, R; k) - \sigma(n, R; k-1) \tag{16}$$

is the number with Drazin index equal to k.

REFERENCES

[1] D. BOLLMAN AND H. RAMIREZ, *On the number of nilpotent matrices over Z_m*, J. Reine Angew. Math., 238 (1969), pp. 85–88.

[2] J. V. BRAWLEY, *The number of nonsingular matrices over a finite ring with identity*. Tech. Rep. 82, Dept. Mathematical Sciences, Clemson Univ., Clemson SC, 1971.

[3] S. L. CAMPBELL AND C. D. MEYER, *Generalized Inverses of Linear Transformations*, Pitman, New York, 1979.

[4] L. E. DICKSON, *Linear Groups with an Exposition of Galois Field Theory*, Dover, New York, 1958.
[5] M. P. DRAZIN, *Pseudoinverses in associative rings and semigroups*, Amer. Math. Monthly, 65 (1958), pp. 506–514.
[6] H. K. FARAHAT, *The multiplicative groups of a ring*, Math. Z., 87 (1965), pp. 378–384.
[7] N. J. FINE AND I. N. HERSTEIN, *The probability that a matrix be nilpotent*, Illinois J. Math., 2 (1958), pp. 499–504.
[8] M. GERSTENHABER, *On the number of nilpotent matrices with coefficients in a finite field*, Illinois J. Math. 5 (1961), pp. 330–333.
[9] R. E. HARTWIG, *Drazin inverses in cryptography*, submitted for publication.
[10] ———, private communication.
[11] ———, A^d *in* $M_n(Z/h)$, 1980, preprint.
[12] J. H. HODGES, *Scalar polynomial equations for matrices over a finite field*, Duke Math. J., 25 (1958), pp. 291–296.
[13] J. LEVINE AND R. E. HARTWIG, *Applications of the Drazin inverse to the Hill cryptographic system*, Cryptologia, 4 (1980), pp. 71–85.
[14] G. Lusztig, *A note on counting nilpotent matrices of fixed rank*, Bull. London Math. Soc., 8 (1976), pp. 77–80.
[15] S. PERLIS, *Theory of Matrices*, Addison-Wesley, Reading, MA, 1952.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 1, March 1982

0196-5212/82/0301-0004 $01.00/0

GENERALIZING THE SUM OF DIGITS FUNCTION*

HELMUT PRODINGER†

Abstract. The number theoretic function $G_{q,\alpha}(n)=\sum_{k\geq 1}\sum_{j=0}^{q-1}\lfloor n/q^k+j\alpha\rfloor$ has appeared in the literature for some special values of α. Some properties of this function are investigated. Since $G_{q,0}(n)$ is closely related to the sum of digits of the q-ary representation of n, a generalized "sum of digits" function can be defined via $G_{q,\alpha}$. For $q=2$ and $\alpha=2^{-s}$ the summing function of this "sum of digits" function is analyzed using a technique of Delange.

1. Introduction and elementary results. Let $q\in\mathbb{N}$, $q\neq 1$ and define the functions $G_{q,\alpha}:\mathbb{N}_0\to\mathbb{N}_0$ by

$$G_{q,\alpha}(n):=\sum_{k\geq 1}\sum_{1\leq j<q}\left\lfloor\frac{n}{q^k}+j\alpha\right\rfloor. \tag{1}$$

($\lfloor x\rfloor$ denotes the greatest integer less or equal to x.)

To make this definition meaningful, α must be in the range $\alpha\in[0,(q-1)^{-1})$. But for all considerations (except for Theorem 5) it is better to restrict α to the range $[0,q^{-1}]$, especially because the generalized "sum of digits" function (see § 2) takes then only nonnegative values, which is very desirable.

In [6] an alternative expression for $G_{2,1/4}$ is given by a complicated method; the same method applies to $G_{2,1/2}$ showing that this function is the identity.

The last result can be found in [4, p. 43] in the general form

$$G_{q,1/q}(n)=\sum_{k\geq 1}\sum_{1\leq j<q}\left\lfloor\frac{n}{q^k}+\frac{j}{q}\right\rfloor=n. \tag{2}$$

In the sequel it will be shown that, starting from (2), some formulas for $G_{q,\alpha}$ can be derived in an easy way. To be able to formulate this result adequately, it is useful to use the following denotation.

If ξ is a string of integers in the range $[0,q-1]$, let $B_q(\xi,n)$ denote the number of subblocks ξ in the q-ary representation of n (subblocks are allowed to overlap).

THEOREM 1.

$$G_{q,q^{-s}}(n)=n-\sum_{1\leq j<q}jB_q(j,n)+\sum_{1\leq j<q}jB_q((q-1)^{s-1}j,n).$$

(For instance, $G_{2,1/4}(n)=n-B_2(1,n)+B_2(11,n)$.)

Proof. It is sufficient to show that the number of indices k, l such that

$$\left\lfloor\frac{n}{q^k}+\frac{l}{q}\right\rfloor=1+\left\lfloor\frac{n}{q^k}+\frac{l}{q^s}\right\rfloor$$

equals

$$\sum_{1\leq j<q}jB_q(j,n)-\sum_{1\leq t<q}tB_q((q-1)^{s-1}t,n).$$

* Received by the editors May 29, 1980, and in revised form June 1, 1981.

† Institut für Algebra und Diskrete Mathematik, Technische Universität Wien, Guszhausstrasze 27-29, A-1040 Vienna, Austria.

This can only happen if

$$\left\lfloor \frac{n}{q^k}+\frac{l}{q} \right\rfloor = 1+\left\lfloor \frac{n}{q^k} \right\rfloor \quad \text{and} \quad s \geqq 2.$$

(For $s=1$ the theorem is trivially fulfilled.) Now assume that the fractional part of n/q^k starts with the digit j, $1 \leqq j < q-1$ (with respect to the q-ary representation of n). Then in (1) each l with $j+l \geqq q$ is possible (there are j of such l's), and furthermore

$$\left\lfloor \frac{n}{q^k}+\frac{l}{q^s} \right\rfloor = \left\lfloor \frac{n}{q^k} \right\rfloor.$$

Now assume $j=q-1$. Then there are again $q-1$ indices in (1) such that $j+l \geqq q$, but

$$\left\lfloor \frac{n}{q^k}+\frac{l}{q^s} \right\rfloor = 1+\left\lfloor \frac{n}{q^k} \right\rfloor$$

is also possible, and this happens if and only if the fractional part of n/q^k starts with $q-1, q-1, \cdots, q-1, t$; in (1) each l with $t+l \geqq q$ is possible (there are t of such l's).

Since the sum over k in (1) means that every digit is exactly one time the leading digit of the fractional part of n/q^k, the proof is finished.

Remark that the formula holds also for $\alpha = 0$ where the second sum vanishes, which can be seen as a "limiting case".

In the sequel it will be shown that $G_{q,\alpha}$ for $0 \leqq \alpha < q^{-1}$ has a rather erratic behavior, which contrasts to the case $\alpha = q^{-1}$.

LEMMA 2. *Assume $a \neq q^{-1}$. Then there exists an n such that*

$$G_{q,\alpha}(n) = G_{q,\alpha}(n+1).$$

Since the proof of this lemma is rather long and not too interesting, we just indicate that an appropriate choice for n is (with respect to the q-ary representation) of the form $(1000\cdots0)_q$.

THEOREM 3. *For $0 \leqq \alpha < q^{-1}$ the function $G_{q,\alpha}$ is not surjective.*

Proof. By Theorem 1,

$$q^t = G_{q,1/q}(q^t) \geqq G_{q,\alpha}(q^t) \geqq G_{q,0}(q^t) = q^t - 1.$$

Thus there are numbers $t_1 < t_2$ such that

$$G_{q,\alpha}(q^{t_2}) - G_{q,\alpha}(q^{t_1}) = q^{t_2} - q^{t_1}.$$

Because of the monotony of $G_{q,\alpha}$, surjectivity in the interval $[t_1, t_2]$ means also injectivity, but this property is not fulfilled.

Remark. If α is allowed to be in the range $\alpha \in [0, (q-1)^{-1})$, $\alpha \neq q-1$ (compare the comments after the definition of $G_{q,\alpha}$), Lemma 2 and Theorem 3 are still true.

It is clear that from $\alpha \leqq \beta$ it follows that $G_{q,\alpha}(n) \leqq G_{q,\beta}(n)$. The following stronger result is easily obtained.

THEOREM 4. *If $\alpha < \beta$ then there is an n such that $G_{q,\alpha}(n) < G_{q,\beta}(n)$.*

Proof. Choose numbers n, k such that

$$1-\beta \leqq \frac{n}{q^k} < 1-\alpha;$$

then

$$\left\lfloor \frac{n}{q^k}+\beta \right\rfloor = 1 \quad \text{and} \quad \left\lfloor \frac{n}{q^k}+\alpha \right\rfloor = 0.$$

As D. E. Knuth has pointed out [5], it would be interesting to investigate $G_{q,\alpha}(n)$ for fixed n, where α is the variable. A first result in this direction is the following theorem.

THEOREM 5.

$$\int_0^1 G_{2,\alpha}(n)\, d\alpha = n.$$

Proof. Since

$$\int_0^1 \lfloor x+\alpha \rfloor\, d\alpha = x,$$

it follows that

$$\int_0^1 G_{2,\alpha}(n)\, d\alpha = \int_0^1 \sum_{k\geqq 1} \left\lfloor \frac{n}{2^k}+\alpha \right\rfloor d\alpha$$

$$= \sum_{k\geqq 1} \int_0^1 \left\lfloor \frac{n}{2^k}+\alpha \right\rfloor d\alpha = \sum_{k\geqq 1} \frac{n}{2^k} = n.$$

(It is not very hard to see that the integration and the summation can be interchanged.)

2. The summing function of the function "generalized sum of digits". In Delange [1] the summing function of the function "sum of digits to the base q" is considered: The sum of digits is

$$\begin{aligned} S_q(n) &= \sum_{r=0}^{\infty} \left(\left\lfloor \frac{n}{q^r} \right\rfloor - q \left\lfloor \frac{n}{q^{r+1}} \right\rfloor \right) \\ &= n-(q-1) \sum_{r=1}^{\infty} \left\lfloor \frac{n}{q^r} \right\rfloor \\ &= n - \sum_{r=0}^{\infty} \sum_{j=1}^{q-1} \left\lfloor \frac{n}{q^{r+1}} \right\rfloor \\ &= n - \sum_{1\leqq j<q} jB_q(j, n) = n - G_{q,0}(n). \end{aligned} \tag{3}$$

In view of § 1 it is natural to define the generalized function "sum of digits" by

$$S_{q,\alpha}(n) := n - G_{q,\alpha}(n).$$

In [1] it is proved that

$$\frac{1}{m} \sum_{n=0}^{m-1} S_q(n) = \frac{q-1}{2} \log_q m + F(\log_q m),$$

where $F(x)$ is continuous, periodic with period 1 and thus bounded. ($\log_q m$ means the logarithm to the base q.) Further information in this area can be found in the beautiful thesis of Flajolet [2].

In the rest of this paper the summing function of $S_{2,2^{-s}}(n)$ is treated, but I hope to do further work in this direction in the future. The ordinary sum of digits appears as the limit for $s \to \infty$.

From Theorem 1, we know that

$$G_{2,2^{-s}}(n) = n - B_2(1, n) + B_2(1^s, n).$$

Plugging this into the definition of $S_{q,\alpha}$, we find that

$$\begin{aligned} S_{2,2^{-s}}(n) &= B_2(1, n) - B_2(1^s, n) \\ &= S_2(n) - B_2(1^s, n). \end{aligned} \tag{4}$$

So $S_{2,2^{-s}}(n)$ is just the number of ones in the binary expansion of n minus the number of blocks of s consecutive ones in that expansion.

The rest of this paper is an analysis of the summing function of the function $B_2(1^s, n)$; then by (4), an analogue to Delange's result of $S_{2,2^{-s}}(n)$ can be formulated as a corollary.

THEOREM 6. *Let $B_2(1^s, n)$ denote the number of subblocks of s consecutive ones appearing in the binary representation of n, where overlapping is allowed. Then the summation of $B_2(1^s, n)$ is given by the formula*

$$\frac{1}{m}\sum_{n=0}^{m-1} B_2(1^s, n) = \frac{\log_2 m - (s-1)}{2^s} + H_s(\log_2 m) + \frac{E}{m},$$

where H_s is continuous, periodic with period 1, and satisfies $H_s(0) = 0$, and where E is bounded by $0 \leqq E < 1$.

Proof. A crucial point in Delange's derivation is the property

$$\left\lfloor \frac{t}{q^r} \right\rfloor = \left\lfloor \frac{n}{q^r} \right\rfloor \quad \text{for } n \leqq t < n+1. \tag{5}$$

For $\alpha \neq 0$ it is not trivial to find an appropriate analogue.

An analogue to property (5) can be written as follows:

$$\left\lfloor \frac{t}{2^r} + \frac{1}{2^s} \right\rfloor = \left\lfloor \frac{n}{2^r} + \frac{1}{2^s} \right\rfloor \tag{6}$$

holds for $n \leqq t < n+1$ and $r \geqq s$ and also for $n - (1/2^{s-r}) \leqq t < n+1-(1/2^{s-r})$ and $r < s$.

Let $l = \log_2 m$. We have

$$\begin{aligned} \sum_{n=0}^{m-1} B_2(1^s, n) &= \sum_{n=0}^{m-1} G_{2,2^{-s}}(n) - \sum_{n=0}^{m-1} G_{2,0}(n) \\ &= \sum_{r=1}^{s-1} \int_{2^{r-s}}^{m-2^{r-s}} \left\lfloor \frac{t}{2^r} + \frac{1}{2^s} \right\rfloor dt + \sum_{r=s}^{\lfloor l \rfloor + 1} \int_0^m \left\lfloor \frac{t}{2^r} + \frac{1}{2^s} \right\rfloor dt - \sum_{r=0}^{\lfloor l \rfloor} \int_0^m \left\lfloor \frac{t}{2^{r+1}} \right\rfloor dt \\ &= -\sum_{r=1}^{s-1} 2^{r-s} \left\lfloor \frac{m}{2^r} + \frac{1}{2^s} \right\rfloor + \sum_{r=0}^{\lfloor l \rfloor} \int_0^m \left(\left\lfloor \frac{t}{2^{r+1}} + \frac{1}{2^s} \right\rfloor - \left\lfloor \frac{t}{2^{r+1}} \right\rfloor \right) dt. \end{aligned}$$

Now define

$$C = \sum_{r=1}^{s-1} 2^{r-s} \left\lfloor \frac{m}{2^r} + \frac{1}{2^s} \right\rfloor,$$

$$g_s(x) = \int_0^x \left(\left\lfloor t + \frac{1}{2^s} \right\rfloor - \lfloor t \rfloor - \frac{1}{2^s} \right) dt.$$

$g_s(x)$ is periodic with period 1, continuous and $g_s(0)=0$. With this notation we can write

$$\sum_{n=0}^{m-1} B_2(1^s, n) = \sum_{r=0}^{\lfloor l\rfloor} \int_0^m \left(\left\lfloor \frac{t}{2^{r+1}}+\frac{1}{2^s}\right\rfloor - \left\lfloor \frac{t}{2^{r+1}}\right\rfloor - \frac{1}{2^s}\right) dt + (\lfloor l\rfloor +1)\frac{m}{2^s} - C$$

$$= \sum_{r=0}^{\lfloor l\rfloor} 2^{r+1} g_s\left(\frac{m}{2^{r+1}}\right) + (\lfloor l\rfloor +1)\frac{m}{2^s} - C$$

$$= \sum_{r=-\infty}^{\lfloor l\rfloor} 2^{r+1} g_s\left(\frac{m}{2^{r+1}}\right) + (\lfloor l\rfloor +1)\frac{m}{2^s} - C$$

$$= \sum_{k=0}^{\infty} 2^{1+\lfloor l\rfloor -k} g_s(m2^{k-\lfloor l\rfloor -1}) + (\lfloor l\rfloor +1)\frac{m}{2^s} - C.$$

Now remember that $m=2^l$ and define $\{l\}=l-\lfloor l\rfloor$ and

$$h_s(x) = \sum_{k\geqq 0} 2^{-k} g_s(x2^k).$$

Then

$$\sum_{n=0}^{m-1} B_2(1^s, n) = m2^{1-\{l\}} \cdot h_s(2^{\{l\}-1}) + (\lfloor l\rfloor +1)\frac{m}{2^s} - C.$$

Now defining

$$H_s(l) = 2^{1-\{l\}} h_s(2^{\{l\}-1}) - \frac{1}{2^s}(\{l\}-1),$$

it remains only to analyze the quantity C to complete the proof.

$$C = \sum_{r=1}^{s-1} 2^{r-s}\left\lfloor \frac{m}{2^r}+\frac{1}{2^s}\right\rfloor = \sum_{r=1}^{s-1} 2^{r-s}\left\lfloor \frac{m}{2^r}\right\rfloor,$$

since r lies in the range $1\leqq r\leqq s-1$. Thus

$$C = \sum_{r=1}^{s-1} 2^{r-s}\left(\frac{m}{2^r}\right) - \sum_{r=1}^{s-1} 2^{r-s}\left\{\frac{m}{2^r}\right\} = m\frac{s-1}{2^s} - E.$$

Since $\{x\}$ lies always in the interval $[0, 1)$, we can deduce that the remaining error term E must also lie in that interval.

Using Delange's result on the summing function of $S_2(n)$ from [1] we get immediately:

COROLLARY 7. *If $S_{q,\alpha}(n)$ denotes the generalized "sum of digits" function defined above, then the summing function of the quantity $S_{2,2^{-s}}(n)$ is given by*

$$\frac{1}{m}\sum_{n=0}^{m-1} S_{2,2^{-s}}(n) = \left(\frac{1}{2}-\frac{1}{2^s}\right)\log_2 m + \frac{s-1}{2^s} + F(\log_2 m) - H_s(\log_2 m) - \frac{E}{m},$$

where both F and H_s are continuous, periodic with period 1, and take the value 0 on the integers, and where E is bounded by $0\leqq E<1$.

3. The Fourier series for $H_s(x)$. Delange [1] has already determined the Fourier series for $F(x)$. Similar methods apply to $H_s(x)$.

THEOREM 8. *The Fourier expansion* $H_s(x)=\sum_k h_k e^{2k\pi ix}$ *of the function* $H_s(x)$ *converges absolutely, and its coefficients are given by*

$$h_0=\log_2\Gamma\left(1-\frac{1}{2^s}\right)-\frac{1}{2^s\log 2}-\frac{1}{2^{s+1}},$$

$$h_k=\frac{\zeta\left(\frac{2k\pi i}{\log 2},1-\frac{1}{2^s}\right)-\zeta\left(\frac{2k\pi i}{\log 2}\right)}{2k\pi i\left(1+\frac{2k\pi i}{\log 2}\right)}\quad \text{for } k\neq 0.$$

Proof. Let $0\leqq x<1$. Since

$$H_s(x)=2^{1-x}h_s(2^{x-1})+\frac{1-x}{2^s},$$

the determination of the Fourier coefficients decomposes as:

$$h_k=\int_0^1 2^{1-x}h_s(2^{x-1})\,e^{-2k\pi ix}\,dx+\frac{1}{2^s}\int_0^1(1-x)\,e^{-2k\pi ix}\,dx=a_k+b_k.$$

It is easily seen that

$$b_k=\frac{1}{2^s}\cdot\frac{1}{2k\pi i}\quad\text{for } k\neq 0,\quad b_0=\frac{1}{2^s}\cdot\frac{1}{2},$$

$$a_k=\int_0^1 2^{1-x}\sum_{r=0}^{\infty}2^{-r}g_s(2^{r+x-1})\,e^{-2k\pi ix}\,dx,$$

and as in [1], the integration and the summation can be interchanged:

$$a_k=\sum_{r=0}^{\infty}\int_0^1 2^{1-r-x}g_s(2^{r+x-1})\,e^{-2k\pi ix}\,dx.$$

The change of variable $x=1-r+\log_2 u$ gives

$$\int_0^1 2^{1-r-x}g_s(2^{r+x-1})\,e^{-2k\pi ix}\,dx=\frac{1}{\log 2}\int_{2^{r-1}}^{2^r}\frac{g_s(u)}{u^2}\exp\left(-2k\pi i\cdot\log_2 u\right)du.$$

Thus

$$a_k=\frac{1}{\log 2}\int_{1/2}^{\infty}\frac{g_s(u)}{u^{2+2k\pi i/\log 2}}\,du.$$

As in [1], the integral

$$\Phi_s(z)=\int_{1/2}^{\infty}\frac{g_s(u)}{u^{z+1}}\,du$$

should be studied; then

$$a_k=\frac{1}{\log 2}\Phi_s\left(1+\frac{2k\pi i}{\log 2}\right).$$

Since

$$g_s(u)=\int_0^u\left(\left\lfloor t+\frac{1}{2^s}\right\rfloor-\lfloor t\rfloor-\frac{1}{2^s}\right)dt,$$

by partial integration for Re $z>0$,

$$\Phi_s(z)=-\frac{1}{2^s}\cdot\frac{2^{z-1}}{z}+\frac{1}{z}\int_{1/2}^{\infty}\left(\left\lfloor u+\frac{1}{2^s}\right\rfloor-\lfloor u\rfloor-\frac{1}{2^s}\right)\frac{du}{u^z}.$$

For Re $z>2$, the integral can be split into three parts. The third one is

$$-\frac{1}{2^s}\cdot\frac{1}{z}\int_{1/2}^{\infty}\frac{du}{u^z}=-\frac{1}{2^s}\cdot\frac{1}{z}\cdot\frac{2^{z-1}}{z-1}.$$

The second one is

$$-\frac{1}{z}\int_{1/2}^{\infty}\lfloor u\rfloor\frac{du}{u^z}=-\frac{1}{z}\cdot\frac{1}{z-1}\cdot\zeta(z-1).$$

The first one is

$$\frac{1}{z}\int_{1/2}^{\infty}\left\lfloor u+\frac{1}{2^s}\right\rfloor\frac{du}{u^z}=\frac{1}{z}\int_{1-2^{-s}}^{\infty}\left\lfloor u+\frac{1}{2^s}\right\rfloor\frac{du}{u^z}=\frac{1}{z(z-1)}\zeta\left(z-1,1-\frac{1}{2^s}\right),$$

where $\zeta(z-1,a)$ is the z-function of Hurwitz (see [3]). This gives

$$\Phi_s(z)=-\frac{1}{2^s}\cdot\frac{2^{z-1}}{z-1}+\frac{\zeta\left(z-1,1-\frac{1}{2^s}\right)-\zeta(z-1)}{z(z-1)}.$$

This holds for Re $z>0$ by analytical continuation. This gives

$$a_k=-\frac{1}{2^s}\cdot\frac{1}{2k\pi i}+\frac{1}{2k\pi i}\left(1+\frac{2k\pi i}{\log 2}\right)^{-1}\left(\zeta\left(\frac{2k\pi i}{\log 2},1-\frac{1}{2^s}\right)-\zeta\left(\frac{2k\pi i}{\log 2}\right)\right)$$

for $k\neq 0$. Now a_0 must be computed. From [7],

$$\zeta(z-1,a)=\tfrac{1}{2}-a+(z-1)(\log\Gamma(a)-\tfrac{1}{2}\log(2\pi))+O((z-1)^2)\quad\text{for } z\to 1.$$

Thus

$$\zeta\left(z-1,1-\frac{1}{2^s}\right)=-\frac{1}{2}+\frac{1}{2^s}+(z-1)\left(\log\Gamma\left(1-\frac{1}{2^s}\right)-\frac{1}{2}\log(2\pi)\right)+O((z-1)^2),$$

$$\zeta(z-1)=\zeta(z-1,1)=-\tfrac{1}{2}-(z-1)\tfrac{1}{2}\log(2\pi)+O((z-1)^2),$$

$$2^{z-1}=1+(\log 2)(z-1)+O((z-1)^2),$$

$$\frac{1}{z}=1-(z-1)+O((z-1)^2).$$

This yields after some manipulations

$$\Phi_s(z)=-\frac{1}{2^s}(1+\log 2)+\log\Gamma\left(1-\frac{1}{2^s}\right)+O((z-1)^2)\quad\text{for } z\to 1.$$

Hence

$$a_0=-\frac{1}{2^s\log 2}-\frac{1}{2^s}+\log_2\Gamma\left(1-\frac{1}{2^s}\right).$$

Finally, since $\zeta(it,a)=O(|t|^{1/2}\log|t|)$ [7], the Fourier series of H_s will converge absolutely.

Acknowledgment. The author would like to express his deep gratitude to an anonymous referee who went very carefully through the manuscript; he pointed out several errors and made numerous suggestions which increase both the readability and the contents of this paper.

REFERENCES

[1] H. DELANGE, *Sur la fonction sommatoire de la fonction somme des chiffres*, l'Enseignement Mathématique, 21 (1975), pp. 31–47.

[2] P. FLAJOLET, *Analyse d'algorithmes de manipulation d'arbres et de fichiers*, Thèse, Université Paris-Sud, 1979.

[3] P. FLAJOLET AND L. RAMSHAW, *A note on Gray code and odd-even merge*, SIAM J. Comput. 9 (1980), pp. 142–158.

[4] D. E. KNUTH, *The Art of Computer Programming*, vol. 1, Addison-Wesley, Reading, MA, 1972.

[5] ———, personal communication.

[6] H. PRODINGER AND F. J. URBANEK, *Infinite* 0-1*-sequences without long adjacent identical blocks*, Disc. Math., 28 (1979), pp. 277–289.

[7] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Cambridge University Press, Cambridge, 1927.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 1, March 1982

0196-5212/82/0301-0005 $01.00/0

MULTIPLYING VECTORS IN BINARY QUADRATIC RESIDUE CODES*

ROBERT CALDERBANK† AND DAVID B. WALES‡

Abstract. Let $q \equiv -1 \pmod 8$ be an odd prime power and let $C(q)$ and $C(q)^*$ be the two extended binary quadratic residue codes of length $(q+1)$. We show how to regard $C(q)$ and $C(q)^*$ as one-sided ideals in a binary group algebra and we show that the appropriate product is a 2-dimensional ideal. This allows us to prove that if d is the minimum weight in $C(q)$ then $(d-1)^2-(d-1)+1-st \geqq q$, where s, t are nonnegative integers, with $s \equiv 0 \pmod 4$, and t odd. The integers s and t depend on the way the nonzero entries of a codeword of minimum weight are distributed among the coordinate positions. We prove that $(d-1)^2-(d-1)+1=q$ only if $q=7$ and $d=4$. We also investigate the case $s=0$.

1. Binary quadratic residue codes. In this section we present properties of binary quadratic residue codes that are needed for the analysis of § 2. We begin by introducing some notation.

We shall denote the vector space $GF(2)^n$ by $V_n(2)$. We shall sometimes denote the zero vector by $\mathbf{0}$ and the all-one vector by $\mathbf{1}$. A binary $[n, k]$ code is a k-dimensional subspace of $V_n(2)$. The weight $\mathrm{wt}(a)$ of a vector $a \in V_n(2)$ is the number of nonzero entries. An $[n, k, d]$ code is a code for which d is the minimum weight among all nonzero codewords. An automorphism of a binary code C is an $n \times n$ permutation matrix P such that $CP = C$. The automorphisms of C form a group which we shall denote by $\mathrm{Aut}(C)$.

Let $q \equiv -1 \pmod 8$ be an odd prime power. If $j \in GF(q)$ then we shall write $j=0$, $j=\square$, or $j \neq \square$ according as j is zero, j is a nonzero square, or j is a nonsquare respectively. The extended (generalized) binary quadratic residue code $C(q)$ is the subspace of $V_{q+1}(2)$ spanned by the rows of the matrix $M(q)$ given below. The rows and columns of $M(q)$ are indexed by the elements of the projective line $GF(q) \cup \{\infty\}$.

$$M(q) = \begin{array}{c|c|ccc|} & \infty & & i & \\ \hline \infty & 1 & 1 & \cdots & 1 \\ \hline i & \begin{matrix}1\\ \vdots \\ 1\end{matrix} & & S & \\ \hline \end{array}$$

where

$$(S)_{ij} = \begin{cases} 1 & \text{if } j-i=\square, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

H. N. Ward [8] and P. Camion [2] define quadratic residue codes in a much more abstract way. J. H. van Lint and F. J. MacWilliams give a simple construction in [3].

* Received by the editors September 19, 1980, and in revised form April 20, 1981. This work was supported in part by the National Science Foundation under contract 50950.

† Bell Laboratories, Murray Hill, New Jersey 07974; formerly at Department of Mathematics, California Institute of Technology, Pasadena, California 91125.

‡ Alfred P. Sloan Laboratory of Mathematics and Physics, California Institute of Technology, Pasadena, California 91125.

Example. The code $C(7)$ is the row space of the matrix

$$M(7)=\begin{array}{c|cccccccc} & \infty & 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline \infty & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 2 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 3 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 4 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 5 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 6 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \end{array}$$

This is the [8, 4, 4] Hamming code. Notice that the incidence matrix of the projective plane $PG(2, 2)$ is a principal submatrix of $M(7)$.

If q is prime then $C(q)$ is a classical quadratic residue code. The matrix S is a circulant and $C(q)$ is an extended cyclic code.

We replace 0 by (-1) throughout $M(q)$ to obtain a $(1, -1)$ matrix $P(q)$. If we regard $P(q)$ as an integral matrix then $P(q)$ is the Hadamard matrix constructed by R. E. A. C. Paley in [6]. Since $q \equiv -1 \pmod 8$ we may write $q+1 = 4t+4$ where t is odd. We have

$$\begin{aligned} SS^T &= ((2t+1)-t)I + tJ \\ &\equiv J \pmod 2, \end{aligned} \tag{2}$$

where J is a square matrix with every entry 1.

It follows that $C(q)$ is self-orthogonal. Note that every row of the matrix $M(q)$ has weight divisible by 4. Since $C(q)$ is self-orthogonal it follows by induction that any sum of rows of $M(q)$ has weight divisible by 4. Thus all weights in $C(q)$ are divisible by 4. (This argument was used by A. M. Gleason, 1962, unpublished).

The group $P\Gamma L(2, q)$ is represented by all permutations of $GF(q) \cup \{\infty\}$ that have the form $(z \to az^\sigma + b/cz^\sigma + d)$, with $a, b, c, d, \in GF(q)$, $ad - bc$ a nonzero square, and σ an automorphism of $GF(q)$. This group is generated by the permutations given as (i), (ii), (iii) and (iv) below.

(i) $T_i = (z \to z + i)$ for $i \in GF(q)$,

(ii) $P_i = (z \to iz)$ for $i = \square$,

(iii) $\rho = (z \to z^p)$ (where p is prime and $q = p^n$), the permutation corresponding to the Frobenius automorphism of $GF(q)$,

(iv) $\tau = (z \to -1/z)$.

We adopt the standard conventions about operations involving ∞. The next result is due to Gleason and Prange.

THEOREM 1.1. *If m_i is the row of $M(q)$ indexed by i then*

(a) $$m_i T_j = m_{i+j} \quad \text{for } j \in GF(q),$$

(b) $$m_i P_j = m_{ij} \quad \text{for } j = \square,$$

(c) $$m_i \rho = m_i P,$$

(d) $$m_i \tau = \begin{cases} m_\infty & \text{if } i = \infty, \\ m_0 + m_\infty & \text{if } i = 0, \\ m_{-1/i} + m_0 & \text{if } i = \square, \\ m_{-1/i} + m_0 + m_\infty & \text{if } i \neq \square. \end{cases}$$

The code $C(q)$ is invariant under $P\Gamma L(2, q)$. (The permutation group acts on the rows of $M(q)$ by multiplication on the right. The permutations are of $GF(q) \cup \infty$.)

Proof. The calculations are relatively routine. The first three are straightforward. The fourth is more difficult as the different cases for i give different results. In each case compare the $(m_i\tau)_j$ coordinate with that on the right-hand side. We omit the details. □

Let λ be the permutation $(z \to -z)$ and let $C(q)^* = C(q)\lambda$. The code $C(q)^*$ is spanned by the vectors $m_i^* = m_{(-i)}\lambda$ for $i \in GF(q) \cup \{\infty\}$. We have

$$m_\infty^* = m_\infty = \mathbf{1},$$

and, if $i \neq \infty$, then

$$(m_i^*)_j = \begin{cases} 1 & \text{if } j = \infty, \\ 1 & \text{if } j - i \neq \square, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Since λ normalizes $P\Gamma L(2, q)$, the code $C(q)^*$ is also invariant under $P\Gamma L(2, q)$. If the coordinates are chosen in the order $\{\infty, 0, \cdots, i, \cdots, -i, \cdots\}$, the code $C(q)^*$ is $C(q)$ with entries other than ∞ and 0 read in reverse.

It is not difficult to show that the codes $C(q)$ and $C(q)^*$ have the properties described in Fig. 1. (Here and elsewhere, $\langle \cdot \rangle$ denotes vector space span.)

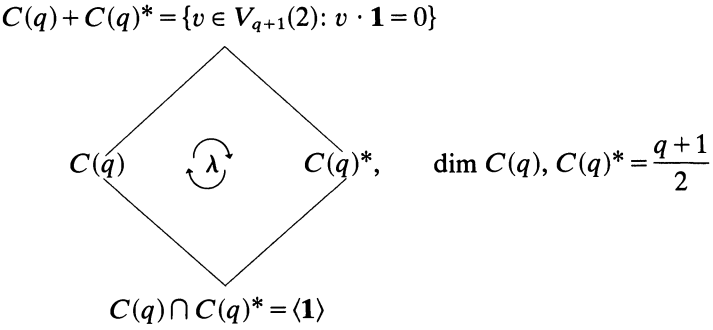


FIG. 1

Let Q and Q^* be the subspaces of $V_q(2)$ obtained from $C(q)$ and $C(q)^*$ respectively by taking each codeword and deleting the entry indexed by ∞. We have $Q^* = Q\lambda$ and $Q^* \cap Q = \langle \mathbf{1} \rangle$. The two codes Q and Q^* are invariant under the group $T = \langle T_i: i \in GF(q) \rangle$. The map

$$\gamma: (\cdots, c_i, \cdots) \to \sum_{i \in GF(q)} c_i T_i$$

is a vector space isomorphism between $V_q(2)$ and the binary group algebra A with basis the permutations in T. Since $(\sum c_i T_i) T_j = \sum c_i T_{i+j}$, we may regard Q and Q^* as ideals in the commutative algebra A. Since the permutations T_i are linearly independent in A, the map $(\sum c_i T_i) \to (\sum c_i)$ from A to $GF(2)$ is well defined. This substitution map is a ring homomorphism.

THEOREM 1.2. *If d is the minimum weight in $C(q)$, then*

$$(d-1)^2 - (d-1) + 1 \geqq q. \tag{4}$$

Proof. Since $P\Gamma L(2, q)$ acts transitively on $GF(q) \cup \{\infty\}$, there exists

$$v = (a_\infty, \cdots, a_i, \cdots) \in C(q),$$

with $a_\infty = 1$ and $\mathrm{wt}(v) = d$. Since $C(q)$ is self-orthogonal and since $\mathbf{1} \in C(q)$ we have $a_\infty = \sum_{i \in GF(q)} a_i$. Now

$$v_1 = (\sum a_i T_i) \in Q \quad \text{and} \quad v_1 \lambda = (\sum a_i T_{-i}) \in Q^*. \tag{5}$$

The product $v_1(v_1\lambda) \in QQ^*$, and since $QQ^* \subseteq Q \cap Q^*$, we have

$$v_1(v_1\lambda) = (\sum a_i T_i)(\sum a_i T_{-i}) = k(\sum T_i), \tag{6}$$

where $k \in GF(2)$. The substitution map gives

$$k = (\sum a_i)^2 = a_\infty^2 = 1.$$

The number of different nonzero terms on the left-hand side of (6) is at most $(d-1)^2 - (d-1) + 1$. This completes the proof (cf. [5]). □

A different proof of Theorem 1.2 is given by J. H. van Lint and F. J. MacWilliams in [3]. H. C. A. van Tilborg [7] generalized Theorem 1.2 as follows.

THEOREM 1.3. *Suppose q is prime. Let d be the minimum weight in $C(q)$. Then*

(a) *If $(d-1)^2 - (d-1) + 1 > q$, then $(d-1)^2 - (d-1) + 1 \geqq q + 12$.*

(b) *If $(d-1)^2 - (d-1) + 1 = q$, then $d = 8t + 4$ and $q = 64t^2 + 40t + 7$ for some t. Furthermore, there exists a projective plane of order $(d-2)$.*

Part (a) of Theorem 1.3 also holds when q is a prime power. The proof rests on analysis of (6) and is essentially the same as that given in [7]. In § 2 we shall prove that equality holds in (4) only if $q = 7$ and $d = 4$. Theorem 1.4 is an intermediate result. It is a special case of a theorem of van Tilborg (see [1]).

THEOREM 1.4. *Suppose that q is a prime power and that the minimum weight d satisfies*

$$(d-1)^2 - (d-1) + 1 = q.$$

Then the vectors of minimum weight in Q are the lines of a projective plane of order $(d-2)$.

Proof. We identify a vector $v \in V_q(2)$ with the set of field elements that index the nonzero entries of v. Equation (6) reveals that the vector $v_1 \in Q$ given by (5) is a difference set in the elementary Abelian group $GF(q)$. The corresponding symmetric block design is the projective plane $PG(2, d-2)$. The lines of this projective plane are the vectors $v_1 T_j$, $j \in GF(q)$.

Let $w \in Q$ be a vector of minimum weight $(d-1)$. Since the code $C(q)$ is self-orthogonal, the vector w meets every line in an odd number of points. There is a line $v_1 T_k$ that meets w in at least 3 points. If there is a point P of $v_1 T_k$ not on w then every other line through P also meets w. But this forces $\mathrm{wt}(w) \geqq (d-2) + 3$, contradicting the choice of w. We conclude that $w = v_1 T_k$, as required. □

2. A stronger form of the square root bound. We may write $v \in V_{q+1}(2)$ in the form

$$v = (a_\infty, \cdots, a_i, \cdots; b_0, \cdots, b_i, \cdots)$$

$$\infty \qquad i \qquad 0 \qquad (-i)$$

nonzero squares; nonsquares

Define

$$d_1(v) = |\{i : a_i \neq 0 \text{ and } i \neq \infty\}|,$$

$$d_2(v) = |\{i : b_i \neq 0 \text{ and } i \neq 0\}|.$$

Since $C(q)$ is self-orthogonal and since $(\mathbf{1};\mathbf{0})$ and $(\mathbf{0};\mathbf{1})$ are vectors in $C(q)$, we have

$$a_\infty = \sum_{i=\square} a_i \quad \text{and} \quad b_0 = \sum_{i=\square} b_i. \tag{7}$$

Let R and R^* be the subspaces of $V_{q-1}(2)$ obtained from $C(q)$ and $C(q)^*$, respectively, by taking each codeword and deleting the entries indexed by ∞ and 0. From Fig. 1 we see that the subspaces R and R^* have the properties described in Fig. 2.

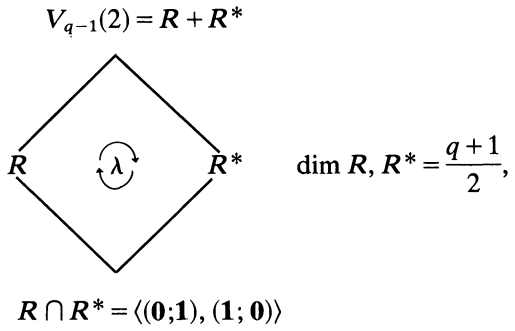


FIG. 2

In § 1 we defined vectors m_k and m_k^* for $k \in GF(q) \cup \{\infty\}$. Let x_k and x_k^* be the vectors obtained from m_k and m_k^* by deleting the entries indexed by ∞ and 0. Then we have $R = \langle x_k : k \in GF(q) \cup \{\infty\} \rangle$ and $R^* = \langle x_k^* : k \in GF(q) \cup \{\infty\} \rangle$.

Since $\tau^2 = 1$, and since $\tau P_i \tau = P_{i^{-1}}$ the group $D = \{P_i, P_i\tau : i = \square\}$ is dihedral of order $(q-1)$. Let B be the binary group algebra with basis the elements of D. Let $\xi : R \to B$ be the linear map given by

$$\xi : (\cdots, a_i, \cdots; \cdots, b_i, \cdots) \to \sum a_i P_i + \sum b_i \tau P_i,$$

and let $\xi^* : R^* \to B$ be the linear map given by

$$\xi^* : (\cdots, c_i, \cdots; \cdots, d_i, \cdots) \to \sum c_i P_i + \sum d_i P_i \tau.$$

Since the elements of D are linearly independent, ξ and ξ^* are injective. We have $\tau P_i = P_{i^{-1}} \tau$ and $\tau(\sum P_i) = (\sum P_i)\tau$, and so ξ and ξ^* agree on $R \cap R^*$. Every element of D acts as an automorphism of both R and R^*. The group D also acts on the algebra B by left and right multiplication. The next lemma connects the different group actions.

LEMMA 2.1. (a) *If $v^* \in R^*$ and $d \in D$, then $\xi^*(v^* d) = d' \xi(v^*)$, where $P_i' = P_i$ and $(P_i\tau)' = \tau P_i$.*

(b) *If $v \in R$ and $d \in D$, then $\xi(vd) = \xi(v)d$.*

(c) *The subspace $\xi(R)$ is a right ideal of B and the subspace $\xi^*(R^*)$ is a left ideal of B.*

Proof. (a) If $v^* = (\cdots, c_i, \cdots; \cdots, d_i, \cdots) \in R^*$, then

$$\begin{aligned} P_j \xi^*(v^*) &= P_j(\sum c_i P_i + \sum d_i P_i \tau) \\ &= \sum c_i P_{ij} + \sum d_i P_{ij} \tau \\ &= \xi^*(v^* P_j) \end{aligned}$$

and

$$\begin{aligned}\tau\xi^*(v^*) &= \tau(\sum c_iP_i + \sum d_iP_i\tau)\\ &= \sum d_iP_{i^{-1}} + \sum c_iP_{i^{-1}}\tau\\ &= \xi^*(v^*\tau).\end{aligned}$$

Part (b) is similar to (a) and we omit the details. Part (c) follows from parts (a) and (b). □

Henceforth, we shall identify R with $\xi(R)$ and R^* with $\xi^*(R^*)$. This allows us to multiply vectors in R^* by vectors in R.

If $x_1 = (\cdots, f_i, \cdots; \cdots, g_i, \cdots)$, then

$$f_i = \begin{cases} 1 & \text{if } i-1 = \square \quad (\text{or if } -i-(-1) \neq \square),\\ 0 & \text{otherwise,}\end{cases}$$

$$g_i = \begin{cases} 1 & \text{if } -i-1 = \square \quad (\text{or if } i-(-1) \neq \square),\\ 0 & \text{otherwise.}\end{cases}$$

From part (c) we have $x^*_{(-1)} = (\cdots, g_i, \cdots; \cdots, f_i, \cdots)$. Identifying R with $\xi(R)$ and R^* with $\xi^*(R^*)$ gives

$$x_1 = \sum f_iP_i + \sum g_i\tau P_i \quad \text{and} \quad x^*_{(-1)} = \sum g_iP_i + \sum f_iP_i\tau.$$

The results in this section rest on the following calculation.

LEMMA 2.2. *If $x_1, x^*_{(-1)}$ are as above, then*

$$(x^*_{(-1)})(x_1) = \tau(\sum P_i).$$

Proof. Since B is a binary algebra, we have

$$\begin{aligned}(x^*_{(-1)})(x_1) &= (\sum g_iP_i + \sum f_iP_i\tau)(\sum f_iP_i + \sum g_i\tau P_i)\\ &= \tau\{(\sum g_iP_{i^{-1}})(\sum g_iP_i) + (\sum f_iP_{i^{-1}})(\sum f_iP_i)\}\\ &= \tau(\sum c_kP_k).\end{aligned}$$

Now we prove that every coefficient $c_k \equiv 1 \pmod 2$.

$$\begin{aligned}c_k = &|\{(i,j): g_i \neq 0, g_j \neq 0 \text{ and } j = ik\}|\\ &+ |\{(i,j): f_i \neq 0, f_j \neq 0 \text{ and } j = ik\}|.\end{aligned}$$

Thus $c_k = (x_1)(x_1P_k)^T$, the inner product of the vectors x_1 and x_1P_k in $V_{q-1}(2)$. Since $x_k = x_1P_k$ we have $c_k = (x_1)(x_k)^T$. If S is the matrix defined in (1), then $(S)_{1,0} = 0$ and (2) gives

$$c_k = \begin{cases} t & \text{if } k \neq 1,\\ 2t+1 & \text{if } k = 1,\end{cases}$$

where $q+1 = 4t+4$. Since t is odd, $c_k \equiv 1 \pmod 2$ as required.

LEMMA 2.3. $R^*R = \langle(\sum P_i), \tau(\sum P_i)\rangle$.

Proof. If $J = \langle(\sum P_i), \tau(\sum P_i)\rangle$, then J is an ideal in B. From Lemma 2.2 we have

$$(x^*_{(-1)})x_1 = \tau(\sum P_i).$$

We multiply on the right by P_k and apply Lemma 2.1 to give

$$(x^*_{(-1)})x_k = \tau(\sum P_i) \tag{8}$$

for all $k = \square$. Multiplying (8) on the right by τ yields

$$(x^*_{(-1)})(x_{-1/k} + x_0) = (\sum P_i)$$

Since $x_0 = (\sum P_i) \in J$, we have

$$(x^*_{(-1)})(x_{-1/k}) \in J \tag{9}$$

for all $k = \square$. Equations (8) and (9) force

$$(x^*_{(-1)})R \subseteq J. \tag{10}$$

Multiplying (10) on the left by τ, and by P_k for all $k = \square$, we obtain $R^*R \subseteq J$ and the proof is complete. □

Since the permutations P_i are linearly independent in B, the map $(\sum c_iP_i) \to (\sum c_i)$ is well defined. This substitution map is a ring homomorphism from the subalgebra of B spanned by the permutations P_i to $GF(2)$.

We define

$$\Omega^* = \{v = (a; b) \in C(q): a_\infty = b_0 = 1 \text{ and } wt(v) = d\},$$

$$\Omega = \{v = (a; b) \in C(q): a_\infty = 1, b_0 = 0 \text{ and } wt(v) = d\}.$$

Then we define

$$s = \max_{v \in \Omega^*} \{|d_1(v) - d_2(v)|\}, \qquad t = \max_{v \in \Omega} \{|d_1(v) - d_2(v)|\}.$$

THEOREM 2.4. *The minimum weight d satisfies*

$$(d-1)^2 - (d-1) + 1 - st \geqq q. \tag{11}$$

Proof. There exists

$$v = (1, \cdots, a_i, \cdots; 0, \cdots, b_i, \cdots) \in \Omega,$$

with $d_1(v) = (((d-1) \pm t)/2)$ and $d_2(v) = (((d-1) \mp t)/2)$. Since $\tau \in \mathrm{Aut}\,(C(q))$ there exists

$$w = (1, \cdots, c_i, \cdots; 1, \cdots, d_i, \cdots) \in \Omega^*,$$

with $d_1(w) = (((d-2) - s)/2)$ and $d_2(w) = (((d-2) + s)/2)$. Now

$$v_1 = \sum a_iP_i + \sum b_i\tau P_i \in R, \qquad w_1 = \sum d_iP_i + \sum c_iP_i\tau \in R^*.$$

Thus $w_1v_1 \in R^*R$, and from Lemma 2.3 we have

$$\begin{aligned} w_1v_1 &= (\sum d_iP_i)(\sum a_iP_i) + (\sum c_iP_i)(\sum b_iP_i) \\ &\quad + \tau\{(\sum c_iP_{i^{-1}})(\sum a_iP_i) + (\sum d_iP_{i^{-1}})(\sum b_iP_i)\} \\ &= k_1(\sum P_i) + k_2\tau(\sum P_i), \end{aligned} \tag{12}$$

where $k_1, k_2 \in GF(2)$. The substitution map gives

$$k_1 = (\sum d_i)(\sum a_i) + (\sum c_i)(\sum b_i),$$

$$k_2 = (\sum c_i)(\sum a_i) + (\sum d_i)(\sum b_i)$$

and from (7) we have $k_1 = 1$ and $k_2 = 1$. Counting nonzero coefficients in (12) using

$k_1 = 1$ or $k_2 = 1$ depending on the sign of $\pm t$ gives

$$\left(\frac{(d-2)-s}{2}\right)\left(\frac{(d-1)+t}{2}\right)+\left(\frac{(d-2)+s}{2}\right)\left(\frac{(d-1)-t}{2}\right)\geqq\frac{(q-1)}{2}.$$

This reduces to (11) and the theorem is proved. □

Remarks. If $v \in \Omega$ then (7) implies that $d_1(v)$ is odd and $d_2(v)$ is even. Thus t is odd and so is not 0. If $v \in \Omega^*$ then $d_1(v)$ and $d_2(v)$ are both odd and $d_1(v)+d_2(v) = d-2$. Since $d \equiv 0 \pmod 4$ we have $d_1(v) \equiv d_2(v) \pmod 4$ and so s is a multiple of 4.

When $q = 7$ we have $d = 4$ and $(d-1)^2-(d-1)+1 = q$. The set Ω^* consists of the three vectors given below:

∞	1	2	4	0	(-1)	(-2)	(-4)
(1	1	0	0;	1	0	0	1)
(1	0	1	0;	1	1	0	0)
(1	0	0	1;	1	0	1	0)

Notice that $s = 0$.

The code $C(23)$ is the [24, 12, 8] Golay code. If $v = (a; b)$ and $w = (c; d)$ where

	∞	1	2	3	4	6	8	9	12	13	16	18	0	(-1)	(-2)	(-3)	(-4)	(-6)	(-8)	(-9)	(-12)	(-13)	(-16)	(-18)
$a =$	(1	0	0	0	1	0	0	0	0	0	0	0)	(0	1	1	1	0	0	0	1	1	1	0	0)
$c =$	(1	0	0	0	1	0	0	0	0	0	0	0)	(1	0	0	0	1	1	1	0	0	0	1	1)

then it is straightforward to check that $v \in \Omega$ and $w \in \Omega^*$, using (41) on [4, p. 498]. We have

$$v_1 = P_4 + \tau(P_1+P_2+P_3+P_9+P_{12}+P_{13}) \in R,$$

$$w_1 = (P_4+P_6+P_8+P_{16}+P_{18})+P_4\tau \in R^*.$$

It is easily checked that $w_1 v_1 = (\sum P_i)+\tau(\sum P_i)$, and in particular we have

$$(\sum P_i) = P_4(P_4+P_6+P_8+P_{16}+P_{18})+P_4(P_1+P_2+P_3+P_9+P_{12}+P_{13}).$$

Notice that $s = 4$, $t = 5$ and $(d-1)^2-(d-1)+1-st = q$.

We now define

$$r = \min_{v\in\Omega}\{|d_1(v)-d_2(v)|\}.$$

THEOREM 2.5. *The minimum weight d satisfies*

$$(d-1)^2-(d-1)+1-((d-3)-r^2)\geqq q. \tag{13}$$

Proof. Let

$$v = (1, \cdots, a_i, \cdots; 0, \cdots, b_i, \cdots) \in \Omega,$$

with $d_1(v) = (((d-1)\pm r)/2)$ and $d_2(v) = (((d-1)\mp r)/2)$. Then

$$v_1 = \sum a_i P_i + \sum b_i \tau P_i \in R,$$

and

$$v_1\lambda = \sum b_i P_i + \sum a_i P_i \tau \in R^*.$$

Since B is a binary algebra and since $(v_1\lambda)v_1 \in R^*R$, we have

$$(14)\qquad \begin{aligned}(v_1\lambda)v_1 &= \tau\{(\textstyle\sum b_iP_{i^{-1}})(\sum b_iP_i)+(\sum a_iP_{i^{-1}})(\sum a_iP_i)\}\\ &= k\tau(\textstyle\sum P_i),\end{aligned}$$

where $k \in GF(2)$. The substitution map gives $k = 1$. Counting different nonzero terms in (14) gives

$$d_1(v)^2 + d_2(v)^2 - (d_1(v)+d_2(v))+1 \geqq \left(\frac{q-1}{2}\right),$$

and so

$$(15)\qquad \left(\frac{(d-1)-r}{2}\right)^2 + \left(\frac{(d-1)+r}{2}\right)^2 - (d-1)+2 \geqq \frac{q+1}{2}.$$

This reduces to (13) and the proof is complete. □

THEOREM 2.6. *If the minimum weight d satisfies $(d-1)^2-(d-1)+1=q$, then $d=4$ and $q=7$.*

Proof. By Theorem 2.4, we have $(d-1)^2-(d-1)+1-st \geqq q$, where $s, t \geqq 0$ and t is odd. This forces $s=0$. By Theorem 1.4 the vectors of minimum weight in the code Q are the lines of a projective plane $PG(2, d-2)$. If $L_1, \cdots, L_{d-1}$ are the lines through 0, then $L_i \cap L_j = \{0\}$ if $i \neq j$. Since $s=0$, every line L_i contains $(d-2)/2$ nonzero squares and $(d-2)/2$ nonsquares. Note that $(d-1)(d-2)/2 = (q-1)/2$ and that each nonzero square is on a unique line L_i. It follows that $P = \langle P_i : i = \square\rangle$ acts transitively on $L_1, \cdots, L_{d-1}$, and that the stabilizer of a line is the subgroup generated by P_γ, where γ is a primitive $(d-2)/2$ root of unity in $GF(q)$. Hence there exist nonzero squares $y_i, z_i \in GF(q)$, $i = 1, \cdots, d-1$, with $y_1 = 1$ such that

$$L_i = \left\{0, y_i\gamma^j, (-z_i)\gamma^j : j = 1, \cdots, \frac{d-2}{2}\right\}$$

for $i = 1, \cdots, (d-1)$. Since the vector $L_1T_{(-1)}$ contains 0 we have $L_1T_{(-1)} = L_m$ for some line L_m. If $d \geqq 8$ then $(d-2)/2 \geqq 3$ and there exist $(\gamma^i - 1), (\gamma^j - 1) \in L_1T_{(-1)}$ such that $(\gamma^j - 1) = \gamma^k(\gamma^i - 1)$, for some $k \neq 0$. If $\alpha = \gamma^j - 1 = \gamma^{k+i} - \gamma^k$ then $L_1T_{(-1)}$ and $L_1T_{(-\gamma^k)}$ are distinct lines in $PG(2, d-2)$ and $0, \alpha \in L_1T_{(-1)} \cap L_1T_{(-\gamma^k)}$. This is impossible, and we conclude that $d=4$. If $d=4$ then (4) implies $q=7$ and the proof is complete. □

As a corollary we have the following theorem on cyclic projective planes.

THEOREM 2.7. *Let $\mathfrak{B}$ be a cyclic projective plane of order $(d-2)$ with $d-2 \equiv 2 \pmod 8$ and with $q=(d-2)^2+(d-2)+1$ a prime.*

If 2 is a primitive $(q-1)/2$ root of unity in $GF(q)$, then $\mathfrak{B} = PG(2, 2)$.

Proof. We note that $q=(d-1)^2-(d-1)+1$ and that $q \equiv -1 \pmod 8$. We may suppose that the points of $\mathfrak{B}$ are labelled with the elements of $GF(q)$ and that $\mathfrak{B}$ is invariant under the cyclic permutation $(z \to z+1)$. If $L = \{e_1, \cdots, e_{d-1}\}$ is a line in $\mathfrak{B}$ then L is a difference set in $GF(q)$. The rows of the incidence matrix of $\mathfrak{B}$ span a cyclic code W. The linear map

$$(a_0, \cdots, a_{q-1}) \to a_0 + a_1x + \cdots + a_{q-1}x^{q-1}$$

allows us to regard W as an ideal in the polynomial ring $K = GF(2)[x]/\langle(x^q-1)\rangle$. Clearly we have $W = \langle(\sum_{i=1}^{d-1} x^{e_i})\rangle$ and

$$\left(\sum_{i=1}^{d-1} x^{e_i}\right)\left(\sum_{i=1}^{d-1} x^{-e_i}\right) = 1 + x + \cdots + x^{q-1}$$

in K. Let α be a primitive qth root of unity in an extension field of $GF(2)$ and let

$$g_0(x) = \prod_{i=\square} (x-\alpha^i) \quad \text{and} \quad g_1(x) = \prod_{i\neq\square} (x-\alpha^i).$$

Since 2 is a primitive $(q-1)/2$ root of unity in $GF(q)$, the polynomials $g_0(x)$ and $g_1(x)$ are irreducible in $GF(2)[x]$. Since $GF(2)[x]$ is a unique factorization domain and since

$$g_0(x)g_1(x) = 1 + x + \cdots + x^{q-1},$$

we have $W = \langle g_0(x)\rangle$ or $W = \langle g_1(x)\rangle$. But it is well known that $Q = \langle g_0(x)\rangle$ and $Q^* = \langle g_1(x)\rangle$. Indeed, this is the way the classical quadratic residue codes are usually defined. The result now follows easily from Theorem 2.6.

This connection between cyclic projective planes and quadratic residue codes was pointed out by van Tilborg in [7].

3. The case s = 0. In view of Theorems 2.4 and 2.5, the possible values of the parameters s, t, and r are of interest. We have seen that s is divisible by 4 and that t, r are both odd. When $s = 0$, Theorem 2.4 gives the same bound on the minimum weight as Theorem 1.2. In this section, we consider a consequence of the condition $s = 0$.

For every vector $v \in \Omega^*$ we define a $d \times d$ matrix $D(v)$ with all entries ± 1. We use the indices of the nonzero entries of v to index the rows and columns of $D(v)$.

$$D(v) = \begin{array}{c|c|c|} & \infty & i \\ \hline \infty & 1 & 1 \cdots 1 \\ \hline i & \begin{matrix}1\\ \vdots\\ 1\end{matrix} & H \\ \hline \end{array},$$

where

$$(H)_{ij} = \begin{cases} 1 & \text{if } j-i = \square, \\ -1 & \text{otherwise.} \end{cases}$$

Notice that $D(v)$ is a principal submatrix of the Paley–Hadamard matrix $P(q)$ that is defined in § 1.

THEOREM 3.1. *If $s = 0$, then for every $v \in \Omega^*$ we have*

$$D(v)D(v)^T = D(v)^T D(v) = dI.$$

Proof. Let d_i be the row of $D(v)$ indexed by i. Since $s = 0$, every vector $x \in \Omega^*$ contains $(d-2)/2$ nonzero squares and $(d-2)/2$ nonsquares. If $i \in v$ and $i \neq \infty$ then $vT_{-i} \in \Omega^*$. The row d_i must have $d/2$ entries 1 and $d/2$ entries (-1), and so $d_i d_\infty^T = 0$.

Let $x \in \Omega^*$, let $\beta \in x$ be a nonzero square and set

$$k_1 = |\{\alpha \in x : \alpha = \square \text{ and } \alpha - \beta = \square\}|,$$

$$k_2 = |\{\alpha \in x : \alpha = \square \text{ and } \alpha - \beta \neq \square\}|,$$

$$k_3 = |\{\alpha \in x : \alpha \neq \square \text{ and } \alpha - \beta = \square\}|,$$

$$k_4 = |\{\alpha \in x : \alpha \neq \square \text{ and } \alpha - \beta \neq \square\}|.$$

Since $x \in \Omega^*$, we have

$$k_1 + k_2 = \left(\frac{d-2}{2}\right) - 1 \quad \text{and} \quad k_3 + k_4 = \left(\frac{d-2}{2}\right). \tag{16}$$

Since $xT_{-\beta} \in \Omega^*$, we have

$$k_1 + k_3 = \left(\frac{d-2}{2}\right) \quad \text{and} \quad k_2 + k_4 = \left(\frac{d-2}{2}\right) - 1. \tag{17}$$

Since $x\tau \in \Omega^*$ and since $-1/\beta \in x\tau$, we have $x\tau T_{1/\beta} \in \Omega^*$. Hence

$$\begin{aligned} \left(\frac{d-2}{2}\right) - 1 &= |\{\alpha \in x : \alpha \neq 0, \infty \text{ and } -1/\alpha + 1/\beta = \square\}| \\ &= \left|\left\{\alpha \in x : \alpha \neq 0, \infty \text{ and } \frac{\alpha - \beta}{\alpha} = \square\right\}\right|, \end{aligned} \tag{18}$$

$$\begin{aligned} \left(\frac{d-2}{2}\right) &= |\{\alpha \in x : \alpha \neq 0, \infty \text{ and } -1/\alpha + 1/\beta \neq \square\}| \\ &= \left|\left\{\alpha \in x : \alpha \neq 0, \infty \text{ and } \frac{\alpha - \beta}{\alpha} \neq \square\right\}\right|. \end{aligned} \tag{19}$$

From (18) we have

$$k_1 + k_4 = \left(\frac{d-2}{2}\right) - 1, \tag{20}$$

and from (19) we have

$$k_2 + k_3 = \left(\frac{d-2}{2}\right). \tag{21}$$

Together (16), (17), (20) and (21) imply

$$k_1 = k_2 = k_4 = \left(\frac{d-4}{4}\right) \quad \text{and} \quad k_3 = \left(\frac{d-4}{4}\right) + 1. \tag{22}$$

Let d_i, $d_j (i, j \neq \infty)$ be two distinct rows of $D(v)$. Without loss we may suppose $i - j = \square$. We have

$$(vT_{-i})T_{(i-j)} = vT_{-j}.$$

If $\alpha \in vT_{-i}$ then $\alpha = k - i$ for some $k \in v$ and if $\alpha' \in vT_{-j}$ then $\alpha' = k' - j$ for some $k' \in v$. If we set $x = vT_{-i}$ and $\beta = j - i$, then (22) implies

$$\begin{aligned} d_i d_j^T &= 1 \cdot 1 + H_{ij}H_{jj} + H_{ii}H_{ji} + k_1 + k_4 - k_2 - k_3 \\ &= 0. \end{aligned}$$

This proves $D(v)D(v)^T = dI$, and since $H^T = -H$, we also have $D(v)^T D(v) = dI$. □

Example. There are 620 codewords of minimum weight 8 in the code $C(31)$. Since the group $\mathrm{PSL}_2(31)$ acts 3-homogeneously on $GF(31) \cup \{\infty\}$, the codewords of weight 8 are the blocks of a 3-design. The seven codewords of weight 8 that contain ∞, 0 and 1

are listed below (The semicolon separates nonzero squares from nonsquares.):

$$v_1 = \{\infty, 1, 9, 25; 0, 6, 26, 27\},$$
$$v_2 = \{\infty, 1, 5, 14; 0, 6, 11, 30\},$$
$$v_3 = \{\infty, 1, 16, 19; 0, 3, 6, 23\},$$
$$v_4 = \{\infty, 1, 2, 7; 0, 6, 12, 15\},$$
$$v_5 = \{\infty, 1, 9, 18; 0, 11, 15, 23\},$$
$$v_6 = \{\infty, 1, 7, 20; 0, 3, 11, 27\},$$
$$v_7 = \{\infty, 1, 5, 25; 0, 15, 3, 13\}.$$

Now

$$\Omega^* = \{v_i P_k;\ k \in GF(31),\ k = \square, \text{ and } i = 1, \cdots, 7\},$$

and we conclude that $s = 0$.

The matrix

$$D(v_1) = \begin{array}{c|cccccccc} & \infty & 1 & 9 & 25 & 0 & 6 & 26 & 27 \\ \hline \infty & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 \\ 9 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 25 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 6 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & -1 \\ 26 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 \\ 27 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \end{array}$$

is an 8×8 Hadamard matrix. If

$$w_1 = \{\infty, 2, 5, 7; 3, 15, 17, 27\},$$
$$w_2 = \{\infty, 4, 5, 7, 9, 19; 17, 29\},$$
$$w_3 = \{\infty, 1, 2, 7, 8, 16; 3, 13\},$$
$$w_4 = \{\infty, 2, 4, 8, 9, 14; 3, 17\},$$

then a similar argument shows that

$$\Omega = \{w_i P_k : k \in GF(31),\ k = \square, \text{ and } i = 1, 2, 3, 4\}.$$

We conclude that $t = 3$ and $r = 1$.

REFERENCES

[1] E. F. ASSMUS, JR., H. F. MATTSON, JR. AND H. E. SACHAR, *A new form of the square-root bound*, SIAM J. Appl. Math., 30 (1976), pp. 352–354.

[2] P. CAMION, *Global quadratic abelian codes*, in Information Theory, C.I.S.M. Courses and Lectures, 219, G. Longo, ed. Springer, Vienna, 1975.

[3] J. H. VAN LINT AND F. J. MACWILLIAMS, *Generalized quadratic residue codes*, IEEE Trans. Inform. theory, IT-24 (1978), pp. 730–737.

[4] F. J. MACWILLIAMS AND N. J. A. SLOANE, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1978.

[5] H. F. MATTSON AND G. SOLOMON, *A new treatment of Bose-Chaudhuri codes*, J. Soc. Ind. Appl. Math., 9 (1961), pp. 654–669.
[6] R. E. A. C. PALEY, *On orthogonal matrices*, J. Math. Phys., 12 (1933), pp. 311–320.
[7] H. C. A. VAN TILBORG, *On weights in codes*, Rep. 71-WSK-03, Dept. of Math., Technol. Univ. of Eindhoven, Netherlands, Dec. 1971.
[8] H. N. WARD, *Quadratic residue codes and symplectic groups*, J. Algebra, 29 (1974), pp. 150–171.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 1, March 1982

0196-5212/82/0301-0006 $01.00/0

THE CLASS OF MEAN RESIDUAL LIVES AND SOME CONSEQUENCES*

MANISH C. BHATTACHARJEE†

Abstract. The class of mean residual life functions and sequences is characterized. Apart from the utility of such characterizations for modelling life distributions through empirically determined mean residual lives, it is shown that such functions arise naturally in many areas such as branching processes. Several additional consequences regarding various nonparametric classes of life distributions are derived, including some characterizations of the exponential and uniform distributions.

1. Introduction and summary. The mean residual life (MRL) of a nonnegative random variable (r.v.) is a well-known concept in reliability theory [2]. In both theory and applications, however, modelling of life distributions has usually been based on the characteristics of the failure rate and the tail of the life distribution. Potentially, the concepts of reliability theory are applicable in studying any phenomenon defined through nonnegative r.v.s, and in this context the mean residual life as a notion of aging has received some attention in the literature in modelling such phenomena as "burn-in" problems of reliability theory [12], [13], duration of strikes and wars, applications to social mobility, labor turnover and staffing policies in organizations [13], [2] (to name a few).

In modelling life distributions of a nonnegative r.v. through empirically determined MRL functions, the first questions is: which functions qualify to be mean residual lives? Although the class of such functions is rich enough, they cannot be entirely freely chosen. In § 2, a complete specification of functions which are MRL functions of some r.v. defined on a subset of $[0, \infty)$ is given which generalizes an earlier incomplete result [11] in this direction. Apart from the obvious utility of defining the permissible set of choices of MRL functions in modelling life distributions, we show that such functions often arise quite naturally in many other areas such as dynamic programming and branching processes. Finally in § 3, a probabilistic interpretation of MRL functions used in the proof of Theorem 2.1 is exploited to yield some new results regarding various classes of life distributions.

For the nonparametric classes of life distributions, we adopt the notations and conventions of Barlow and Proschan [2]. In particular, the concepts IFR, DFR, IMRL, DMRL, NBU, NBUE, NWU, NWUE are defined and discussed there.

2. Characterization of MRL functions and examples. For a cumulative distribution function (cdf) $F \in L^1$ on $[0, \infty)$ of a r.v., μ_F generically denotes the mean, $\bar{F} = 1 - F$ the tail (reliability), $r_F(x)$ the failure rate (when it exists), $g_F(x)$ the mean residual life (MRL), and we let

$$x^*(F) = \sup \{t > 0 : F(t) < 1\} \leqq \infty$$

be the upper endpoint of the support of F.

2.1. We ask the following question: given a real-valued $g : (0, x_0) \to (0, \infty)$, for some $x_0 \in (0, \infty]$, when does there exist a nondegenerate r.v. $X \geqq 0$ a.s. such that g

* Received by the editors January 28, 1980, and in final form April 20, 1981. This work was done at the Indian Institute of Management Calcutta, and was supported in part by the National Research Council of Canada under grant NRC A-8218.

† Indian Institute of Management Calcutta, Calcutta 700 027, India, and Laurentian University, Sudbury, Ontario, Canada P3E-2C6.

has a representation

$$g(t) = E(X - t | X > t)? \tag{2.1}$$

If such an X exists, is it unique?

Given the cdf of a r.v. $X \in L^1$, its MRL is

$$E(X - t | X > t) \equiv g_F(t) = \int_t^{x^*(F)} \frac{\bar{F}(x)\, dx}{\bar{F}(t)}, \qquad 0 < t < x^*(F); \tag{2.2}$$

we let TF denote the induced distribution

$$TF(t) = \int_0^t \left(\frac{\bar{F}(x)}{\mu_F} \right) dx, \qquad 0 < t < x^*(F). \tag{2.2a}$$

THEOREM 2.1 (cf. [11, Thm. 2]). *Let $g: (0, x_0) \to (0, \infty)$, for some $0 < x_0 \leqq \infty$. Then g is a* MRL *function if and only if g is right-continuous, $0 < g(0+) < \infty$, $\int_0^{x_0} dx/g(x) = \infty$ and*

$$J(t) = \int_0^t \frac{dx}{g(x)} + \log g(t) \text{ is nondecreasing } (\uparrow) \text{ on } (0, x_0). \tag{2.3}$$

The cdf F for which $g = g_F$ is a.s. unique with $x^*(F) = x_0$.

Proof. The key idea is the following observation which is implicit in a paper of Meilijson [7]. If F is a nondegenerate cdf with MRL g, then (2.2) and (2.2a) imply that the reciprocal of g is the failure rate [2] of TF, so that

$$\overline{TF}(t) = \exp\left(-\int_0^t \frac{dx}{g(x)} \right), \qquad 0 < t < x^*(F) \tag{2.4}$$

by a well-known formula in reliability theory; a fortiori

$$\bar{F}(t) = \frac{\mu}{g(t)} \exp\left(-\int_0^t \frac{dx}{g(x)} \right), \qquad 0 < t < x^*(F), \tag{2.5}$$

setting up (with (2.2)) a 1:1 correspondence between a cdf and its MRL. The theorem is now obvious. For the necessity of $0 < g(0+) < \infty$, note $g(0+) = \mu_F / \bar{F}(0+) \in (0, \infty)$ since μ_F does. □

An incomplete answer to the question in (2.1), restricted to strictly positive r.v.s with infinite support and absolutely continuous distributions was given earlier by Swarz [11]. It is weaker than our theorem; he makes the unnecessary assumption that g is strictly positive and differentiable ($g'(t) \geqq -1$) on all of $(0, \infty)$. The probabilistic interpretation $g_F = 1/r_F$ used in our arguments will later yield other useful implications. As a consequence of (2.2) and the representation (2.5), we also have the immediate but useful conclusion:

COROLLARY 2.1. *Let F_n be a sequence of cdfs satisfying: $\bar{F}_n(x) \leqq B(x)$, $x > 0$, $n \geqq 1$, for some integrable $B(x)$. Then F_n converges weakly to some cdf F if and only if $g_{F_n}(x) \to g_F(x)$ at every continuity point $x < x^*(F)$ of F (and hence of g_F).*

The proof uses the dominated convergence theorem; the *if* part follows by contradiction using Helly's selection theorem and the 1:1 correspondence between cdfs and their MRLs as noted in Theorem 2.1. For distributions on $(-\infty, \infty)$, a parallel recent result of Kotz and Shanbhag [14] requires an additional condition on the asymptotic negligibility of left tails.

Proof. Assume that F_n converges weakly to some cdf F. Taking any x with $0 < x < x^*(F)$, we can find an integer $n_0(x)$ such that $n \geqq n_0(x)$ implies $\bar{F}_n(x) > 0$. The

corresponding MRLs

$$g_{F_n}(x) = \int_x^\infty \frac{\bar{F}_n(y)\,dy}{\bar{F}_n(x)}, \qquad n \geqq n_0(x)$$

exist and are finite for every such x, since as argued the denominator is positive, while the given condition that $\bar{F}_n$ be bounded above by an integrable function B implies that each F_n is in L^1, so that the numerator above is also finite. In virtue of the same condition, the Lebesgue dominated convergence theorem applies to the numerator above. If $0 < x$ $(< x^*(F))$ is also a continuity point of F, the denominator converges too and we get $g_{F_n}(x) \to \int_x^\infty \bar{F}(y)\,dy/\bar{F}(x) = g_F(x)$.

Conversely, suppose if possible that F_n does not converge weakly to F but $g_{F_n} \to g_F$ at every continuity point of F. By Helly's first theorem, there exists a subsequence F_{n_i} converging to some nondecreasing right-continuous F^* different from F. Again appealing to the dominated convergence theorem and the hypothesis that g_{F_n} converge, we get for all x such that $0 < x < x^*(F) \wedge x^*(F^*)$,

$$g_F(x) = \lim_{i\to\infty} g_{F_{n_i}}(x) = \int_x^\infty \frac{\bar{F}^*(y)\,dy}{\bar{F}^*(x)}, \tag{2.5a}$$

where $\wedge$ stands for minimum. Note F^* is a proper cdf on $[0, \infty)$ with a possible jump at 0, that is, $1 - F_n \leqq B$ integrable implies

$$1 - \lim_{x\to\infty} F^*(x) = \lim_{x\to\infty} \limsup_{i\to\infty} (1 - F_{n_i}(x)) \leqq \lim_{x\to\infty} B(x) = 0$$

since $\int_0^\infty B(y)\,dy < \infty$. Accordingly, F^* is uniquely specified by its MRL g_{F^*} (viz. Theorem 2.1). But $g_{F^*} = g_F$ by (2.5a) above, provided $x^*(F) = x^*(F^*)$. We can then conclude that $F = F^*$, a contradiction.

To complete the proof, it remains to confirm that $x^*(F) = x^*(F^*)$. First note [14] that using (2.5a) gives

$$x^*(F) = \inf\left\{t: \lim_{x\to t} g_F(x) = 0\right\} \leqq \inf\left\{t: \lim_{x\to t} g_{F^*}(x) = 0\right\} = x^*(F^*),$$

the inequality above arising from the fact that $g_F = g_{F^*}$ on $(0, x^*(F) \wedge x^*(F^*)) \subset (0, x^*(F))$. Hence suppose $x^*(F) < x^*(F^*)$, if possible. Then, using (2.5a) again, we get

$$\infty = \int_0^{x^*(F)} \frac{dy}{g_F(y)} = \int_0^{x^*(F)} \frac{dy}{g_{F^*}(y)} < \int_0^{x^*(F^*)} \frac{dy}{g_{F^*}(y)} = \infty,$$

a contradiction. The only remaining possibility is $x^*(F^*) = x^*(F)$. □

2.2. Theorem 2.1 identifies the set of MRL functions from which a model builder may choose. We now give some interesting examples of MRL functions. I do not know of any direct application of the following examples, but they are of independent interest and serve to illustrate the richness of the class of MRL functions by showing how they arise naturally in other contexts.

Examples. (i) Any right-continuous nonincreasing function g on $(0, \infty)$, satisfying (2.3), is a MRL. This is obvious if we choose $\varepsilon > 0$ and note that

$$\int_0^\infty \frac{dx}{g(x)} \geqq \int_\varepsilon^\infty \frac{dx}{g(\varepsilon)} = \infty.$$

(ii) For any failure rate function r on $(0, \infty)$ and $\downarrow$ with $0 < r(0+) < \infty$, $[r(x)]^{-1}$ is a MRL. For any nondiscrete failure distribution on $(0, \infty)$ with a density, this follows

from the well-known representation [2]

$$\bar{F}(t) = \exp\left(-\int_0^t r(x)\,dx\right), \qquad 0 < t < x^*(F).$$

(iii) Let $M(t)$ be the renewal function generated by a "new better than used in expectation" (NBUE) cdf F. Then for any $\varepsilon > 0$ with $F(\varepsilon) > 0$, $M(t+\varepsilon)$ is a MRL. To prove the assertion, it suffices to note that F is NBUE $\Rightarrow M(t) \leqq t/\mu_F$ [2, p. 171]. Hence

$$\int_0^\infty \frac{dx}{M(x+\varepsilon)} \geqq \mu_F \int_0^\infty \frac{dx}{x+\varepsilon} = \infty.$$

Thus, since $\sum_{k=1}^\infty F^{(k)}(t+\varepsilon)$ is a MRL ($F^{(k)}$ denotes the kth convolution of F), by letting $\varepsilon \to 0$ we see that for a NBUE distribution with 0 as left endpoint of support, there exists a sequence of r.v.s X_n such that

$$M(t) = \lim_{n\to\infty} E(X_n - t | X_n > t).$$

$M(t)$ itself is not a MRL since $M(0) = 0$.

(iv) Optimal disposal of an asset (variant of a problem treated by Karlin [6]). An asset for which a sequence of i.i.d. price (Y) quotations are sequentially received can be sold any day at the best price offered so far, while it costs $c > 0$ to maintain it each day the asset is kept by rejecting offers. If P_n is the optimal gain for n-day horizon, the limiting optimal gain $P^* = \lim P_n$ is finite, being the only root of the equation $A(x) = 0$, where

$$A(x) = x - E \max(Y, x - c) = E \min(c, x - Y),$$

and further it can be shown following Karlin [6] that

$$P^* = \min\left\{a : \int_0^a \frac{dx}{A(x)} = \infty\right\}, \qquad 0 \leqq A(x) \text{ is } \uparrow.$$

Thus $A(x)$ is a MRL of a distribution supported by $(0, P^*)$.

(v) Let $f(s) = Es^Z$, $0 < s < 1$ with $P(Z = 1) \neq 1$, be the probability generating function (pgf) of the progeny Z of a critical/subcritical ($EZ \leqq 1$) continuous time Markov branching process $\{Z(t) : t > 0\}$ satisfying the "nonexplosion hypothesis" $P\{Z(t) < \infty\} = 1$, for all $t > 0$. Then there exists an a.s. unique r.v. X on the unit interval such that

$$f(s) = E(X | X > s).$$

To verify this, note that the condition that $P(Z = 1) \neq 1$, $EZ \leqq 1 \Rightarrow f(s) > s$ on $(0, 1)$ and the nonexplosion hypothesis holds [1], is equivalent to the statement

$$\forall \varepsilon > 0, \quad \int_{1-\varepsilon}^1 \frac{ds}{f(s) - s} = \infty \Rightarrow \int_0^1 \frac{ds}{f(s) - s} = \infty.$$

Since

$$J(s) = \int_0^s \frac{dx}{f(x) - x} + \log(f(s) - s) \Rightarrow J'(s) = \frac{f'(s)}{f(s) - s} > 0, \qquad 0 < s < 1,$$

(2.3) holds. Since f is continuous on $(0, 1)$ and

$$EZ \leqq 1 \Rightarrow \lim_{s \to 0+} [f(s) - s] = f(0) = P(Z = 0) > 0,$$

$f(s) - s$ is a MRL on $(0, 1)$, from which the claim follows.

(vi) Let $\{Z(t): t > 0\}$ be a Markov branching process with the usual infinitesimal generating function $u(s) = \sum_{k=0}^{\infty} a_k s^k$ satisfying $u'(1) < 0$ (i.e., almost sure extinction). Then it is well known [1] that, conditioned to nonextinction, $Z(t)$ has a proper limit distribution

$$\lim_{t \to \infty} P(Z(t) = n | Z(t) > 0) = b_n$$

whose pgf $B(s)$ is determined by

$$B(s) = 1 - \exp\left\{u'(1) \int_0^s \frac{dx}{u(x)}\right\}, \qquad 0 < s < 1.$$

Let $g(s) = -u(s)/u'(1)$, $0 < s < 1$. Since $B(s) \to 1$ as $s \to 1-$, $1/g$ is a failure rate but g is *not* a MRL unless $u(s)$ is linear. The argument is as follows. Since $u'(s) \uparrow$ by the convexity of $u(s)$, if $u(s)$ is not linear then we must have $u'(s) \leqq u'(1)$ with strict inequality for some s_0 *in* $(0, 1)$, since $u'(s)$ is not constant. This violates (2.3), which requires $u'(s) \geqq u'(1)$ for all s in $(0, 1)$. In the remaining case (fractional linear generating function of offspring), g is a MRL.

(vii) Consider any critical Galton–Watson process with infinite mean time (T) to extinction and offspring pgf $f(s) = E(s^Z)$ satisfying $\operatorname{var}(Z) \leqq EZ$. Then (a) $[f(s) - s]/(1 - s)$ is a MRL; (b) the tail of the offspring distribution is the first passage time distribution of a renewal sequence. Since for a critical Galton–Watson process,

$$ET < \infty \Leftrightarrow \int_0^1 \frac{1 - s}{f(s) - s}\, ds < \infty \qquad \text{(Seneta [9])}, \tag{2.6}$$

given $ET < \infty$, a) follows provided

$$J(s) = \int_0^s \frac{1 - x}{f(x) - x}\, dx + \log(f(s) - s) - \log(1 - s)$$

is nondecreasing. Compute

$$J'(s) = \frac{f'(s) - s}{f(s) - s} + \frac{1}{1 - s}.$$

We show $J'(s) \geqq 0$ on the unit interval. Note that the offspring pgf $f(s)$ is critical $(f'(1) = 1)$ and $\operatorname{var} Z \leqq EZ$ implies $f''(1) = EZ(Z - 1) = EZ^2 - 1 = \operatorname{var} Z \leqq EZ$. This implies that $f'(s)$ is a subcritical pgf. Hence $f'(s) > s$ on $(0, 1)$. Thus $J(s)$ is nondecreasing.

The proof of the remaining claim uses the MRL property of $(f(s) - s)/(1 - s)$ under the stated conditions. To assert (b), note the reciprocal of this MRL can be expressed as

$$U(s) \overset{\text{def}}{=} \frac{1 - s}{f(s) - s} = \frac{1}{1 - H(s)}, \qquad 0 < s < 1 \tag{2.7}$$

where $H(s)=[1-f(s)]/1-s$ is the generating function of the tail $h_n \overset{\text{def}}{=} P(Z \geqq n)$ of the progeny Z, $n=1, 2, \cdots$. Now the representation (2.7) shows that $\{h_n: n \geqq 1\}$ is the first passage time distribution of a renewal sequence $\{u_n: n \geqq 0\}$ with generating function $U(s)$ such that $[U(s)]^{-1}$ is a MRL on $(0, 1)$. Since, given $ET=\infty$, (2.6) requires

$$\infty = \int_0^1 U(s)\, ds = \sum_{n=0}^{\infty} \frac{u_n}{n+1} \leqq \sum_{n=0}^{\infty} u_n = \left(1 - \sum_{n=1}^{\infty} h_n\right)^{-1},$$

the first passage time distribution is "honest" (i.e., no atom at ∞).

(viii) Any slowly varying function L with $0<L(0+)<\infty$ and $\uparrow$ is a MRL. This is true since, if L is slowly varying [4], then (a) so is $1/L(x)$ and (b) $\int_0^{\infty} L(x)\, dx = \infty$.

As an application, consider a supercritical Galton–Watson process $\{Z_n: n = 0, 1, 2, \cdots\}$ with a single ancestor ($Z_0 = 1$ a.s.) and mean progeny $EZ_1 = m > 1$. Then we know that under fairly general conditions (i.e., such that $EZ_1 \log Z_1 < \infty$), (Z_n/m^n) converges in distribution to a nondegenerate r.v. W with $EW = 1$; it is also true [1] that $\int_0^t P(W>x)\, dx$ is slowly varying. Hence, for any $\varepsilon > 0$,

$$L(t, \varepsilon) = \varepsilon + \int_0^t P(W>x)\, dx, \qquad t>0 \tag{2.8}$$

is also slowly varying, satisfies the conditions in (viii) and is thus a MRL function. Reparametrizing with $\varepsilon = 1/n$, Corollary 2.1 implies

$$\int_0^t P(W>x)\, dx = \lim_{n\to\infty} E(X_n - t | X_n > t)$$

for some sequence X_n with decreasing mean residual lives (DMRL). To check the DMRL nature of X_n note that, since W has a nonvanishing density on $(0, \infty)$, (2.8) implies that the MRL $L(t, n^{-1}) = E(X_n - t | X_n > t)$ is convex nonincreasing in t for each n.

For MRLs defined on a finite interval, a suitable reparametrization will generally yield a corresponding MRL in $\mathcal{G}_\infty$ (the set of MRLs with support $(0, \infty)$). For example, for any critical pgf f with $ET = \infty$ as in example (vii),

$$e^{2t}[f(1-e^{-t}) - (1-e^{-t})] \in \mathcal{G}_\infty,$$

being the MRL of the r.v. $G_0^{-1}(Y)$, where Y on $(0, 1)$ has the MRL of example (vii) and G_0 is the exponential cdf with mean one.

3. Applications. The induced distribution TF, whose probabilistic interpretations in the context of renewal theory are well known, its successive iterates T^nF and the relationship $r_{TF} = 1/g_F$ yield the following results.

THEOREM 3.1. i) *Suppose $F \in L^2$. Then F and TF have the same mean residual life if and only if F is exponential.*

ii) *Let F be* NBU. *Then $\mu_{TF} = \mu_F$ if and only if F is exponential.*

Proof. i) Necessity follows from the fact that $TF = F$ if and only if F is exponential. Conversely suppose $g_{TF} = g_F$. Now $F \in L^2$ guarantees the existence of T^2F. Note $1/g_{TF}(x)$ is the failure rate of $T^2F = TG$ where $G \overset{\text{def}}{=} TF$, while $1/g_F(x)$ is the failure rate of G. Hence

$$g_{TF} = g_F \Leftrightarrow TG = G \Leftrightarrow G, \quad \text{i.e., } TF \text{ is exponential}$$
$$\Leftrightarrow F \text{ is exponential.}$$

ii) We only need to prove necessity. Since F is NBU, i.e., $\bar{F}(x+y) \leqq \bar{F}(x)\bar{F}(y)$ for all $x, y > 0$, we have

$$0 = \mu_F - \mu_{TF} = \int_0^\infty \bar{F}(x)\,dx - \int_0^\infty \overline{TF}(x)\,dx$$

$$= \mu_F^{-1} \int_0^\infty \left[\mu_F \bar{F}(x) - \int_0^\infty \bar{F}(x+y)\,dy\right] dx$$

$$= \mu_F^{-1} \int_0^\infty \int_0^\infty [\bar{F}(x)\bar{F}(y) - \bar{F}(x+y)]\,dy\,dx.$$

Thus the integrand, being nonnegative, must vanish almost everywhere, implying exponentiality of F. □

The same condition $\mu_F = \mu_{TF}$ characterizes the exponential within the NWU class under the additional assumption $F \in L^2$, which is free for NBU distributions ([2] and Theorem 3.3ii) below).

THEOREM 3.2. *Let Y, X_1, X_2 be independent nonnegative r.v.s such that X_1, X_2, are i.i.d. with a pdf and $EX_1 < \infty$. Then*

$$|Y - X_1| \overset{d}{=} Y \overset{d}{=} \min(X_1, X_2) \tag{3.1}$$

if and only if X_1 is uniformly distributed ($\overset{d}{=}$ *denotes equality in distribution*).

Proof. Let F be the cdf of X_1. Then, under the stated conditions, $Y \overset{d}{=} |Y - X_1|$ if and only if Y is absolutely continuous and distributed as TF [8, Thm. 3]. Hence (3.1) is equivalent to

$$\int_t^{x^*(F)} \bar{F}(x)\,dx = C[\bar{F}(t)]^2, \qquad 0 < t < x^*(F), \tag{3.2}$$

where $C = \mu_F$. If f is the density of F, this implies $\bar{F}(t) = 2C\bar{F}(t)f(t)$, or,

$$f(t) = (2C)^{-1}, \qquad 0 < t < x^*(F).$$

Since f is a probability density, this requires $x^*(F) = 2C < \infty$. The converse follows by direct computation by showing (3.2) holds when F is uniform. □

THEOREM 3.3. *Let F be a nondiscrete failure distribution. Then*:

i) *F is* DMRL $\Leftrightarrow$ *TF is IFR $\Leftrightarrow$ T^2F is strongly unimodal.*

ii) *Stieltjes' moment problem is determined for* NBU *distributions F.*

iii) *If F is* DFR *with $f(0+) < \infty$, then*

$$\bar{F}(t) \geqq \bar{F}(0+) \exp\left\{-\frac{xf(0+)}{\bar{F}(0+)}\right\}, \qquad x > 0. \tag{3.3}$$

Equality holds if and only if F is exponential.

Proof. i) Note a) a probability density f is "strongly unimodal" (i.e., its convolution with any unimodal distribution is unimodal) if and only if f is log-concave on $\{x: f(x) > 0\}$ and there are no gaps in the interval of support (Ibragimov [5]). Also (2.3) implies b) F is DMRL if and only if TF is IFR and c) F is IFR if and only if TF is strongly unimodal. Combining a), b) and c) yields the claim.

ii) The existence of all moments of a NBU distribution is an easy consequence of the fact that F is NBU $\Rightarrow \log(\mu_n/n!)$ is subadditive [2, p. 187]. The following alternative argument is instructive. We first argue that if NBU $F \in L^1$, then F must

have all moments and subsequently verify that $F \in L^1$. By the NBU property of F,

$$\overline{TF}(x) = \mu_F^{-1} \int_0^\infty \bar{F}(x+y)\, dy \leqq \mu_F^{-1} \int_0^\infty \bar{F}(x)\bar{F}(y)\, dy = \bar{F}(x).$$

Thus $TF \leqq^{\text{st}} F$, where $\leqq^{\text{st}}$ denotes stochastic majorization. Hence $\mu_{TF} = \int_0^\infty \overline{TF}(x)\, dx \leqq \mu_F < \infty$, i.e., $TF \in L^1$. But $TF \in L^1$ if and only if $F \in L^2$. Hence T^2F exists. Repeating the argument, we get

$$F \overset{\text{st}}{\geqq} TF \overset{\text{st}}{\geqq} T^2F \overset{\text{st}}{\geqq} \cdots \overset{\text{st}}{\geqq} T^nF \overset{\text{st}}{\geqq} \cdots.$$

We thus have a stochastically decreasing sequence $T^nF \in L^1$. But $T^nF \in L^1$ if and only if $F \in L^{n+1}$, $n = 1, 2, \cdots$. So given $F \in L^1$, $F \in L^p$ for all $p \geqq 1$. To check $F \in L^1$, choose $x_0 \in (0, x^*(F))$ with $\bar{F}(x_0) < 1$. Then

$$\begin{aligned}\mu_F = \sum_{n=0}^\infty \int_{[nx_0,(n+1)x_0)} \bar{F}(x)\, dx &\leqq \sum_{n=0}^\infty x_0 \bar{F}(nx_0) \\ &\leqq x_0 \sum_{n=0}^\infty [\bar{F}(x_0)]^n \quad \text{by the NBU property} \\ &= \frac{x_0}{F(x_0)} < \infty.\end{aligned}$$

Given that F is NBU, its moment sequence $\{\mu_n\}$ satisfies $\mu_n \leqq n!\mu_1^n$, $n = 1, 2, \cdots$ (see [2]). Hence

$$\sum_{n=0}^\infty \mu_{2n}^{-1/2n} \geqq \sum_n \{(2n)!\mu_1^{2n}\}^{-1/2n} \geqq \mu_F \sum_n \{(2n)^{2n}\}^{-1/2n} = \frac{1}{2}\mu_F \sum_n \frac{1}{n} = \infty.$$

Thus the "Carleman condition" [4] holds, and accordingly $\{\mu_n\}$ uniquely determines F.

iii) Since F is DFR, given $f(0+) < \infty$,

$$\bar{G}(x) = \frac{f(x)}{f(0+)}, \qquad x > 0 \tag{3.4}$$

defines the tail of a distribution G continuous on $[0, \infty)$ (viz. [2], F is DFR$\Rightarrow F$ has a strictly positive $\downarrow$ density on $(0, \infty)$) with mean $\mu_G = \bar{F}(0+)/f(0+)$. Hence

$$\overline{TG}(x) = \mu_G^{-1} \int_x^\infty \bar{G}(t)\, dt = \frac{\bar{F}(x)}{\bar{F}(0+)}, \qquad x > 0. \tag{3.5}$$

From (3.4), the mean residual life of G is $g_G(x) = 1/r_F(x)$ is increasing on $(0, \infty)$. Thus G is IMRL $\subset$ NWUE, so that a well-known inequality for NWUE distributions [2, p. 187] yields

$$\mu_G \overline{TG}(x) = \int_x^\infty \bar{G}(t)\, dt \geqq \int_x^\infty \exp\left(-\frac{t}{\mu_G}\right) dt = \mu_G \exp\left(-\frac{x}{\mu_G}\right), \qquad x > 0. \tag{3.6}$$

Combining (3.5) and (3.6), the result follows.

If F is exponential, equality holds in (3.3). Conversely, if equality holds in (3.3), then $f(x) = \bar{F}(x)f(0+)/\bar{F}(0+)$ on $(0, \infty)$ by differentiation, i.e., $r_F(x)$ is constant. □

Remark 1. The bound (3.3) is of course nontrivial only if $r_F(0+) < \infty$. 2. Since (2.4) and (2.5) show that $\bar{F}(x) = (\mu_F/g_F(x))\overline{TF}(x)$, it follows that F is NBUE$\Leftrightarrow$ $TF \leqq^{\text{st}} F$. The argument in the proof of Theorem 3.2ii) then implies that all moments

of NBUE distributions are finite. 3. Note F is IMRL$\Rightarrow g_F(x) \geqq g_F(0+) = \mu_F \Rightarrow F \leqq^{st} TF \Leftrightarrow F$ is NWUE, as used in (3.6).

The following property of MRL functions is dual to the closure property of failure rate functions under addition (resulting from formation of series systems).

PROPOSITION. *Let $g_1, g_2 \in \mathcal{G}_\infty$. Then there is a $g \in \mathcal{G}_\infty$ such that*

$$\frac{1}{g} = \frac{1}{g_1} + \frac{1}{g_2} \tag{3.7}$$

Proof. Let $g_i(t)$ be the MRL of F_i with mean μ_i $(i = 1, 2)$. Consider the cdf

$$G(t) = 1 - \frac{\mu_2 \bar{F}_1(t)\overline{TF}_2(t) + \mu_1 \overline{TF}_1(t)\bar{F}_2(t)}{\mu_2 \bar{F}_1(0+) + \mu_1 \bar{F}_2(0+)}.$$

Then the MRL of G satisfies (3.7). If neither F_i $(i = 1, 2)$ has an atom at 0, then

$$\bar{G}(t) = \alpha \bar{F}_1(t)\overline{TF}_2(t) + (1-\alpha)\overline{TF}_1(t)\bar{F}_2(t),$$

represents the reliability of a mixture of two series systems with elements (F_1, TF_2) and (TF_1, F_2) respectively, and where $\alpha = \mu_2/(\mu_1 + \mu_2)$. □

We close with the discrete analogue of Theorem 2.1. A positive sequence $\{\lambda_n, n = 0, 1, 2, \cdots\}$ is a MRL *sequence* if there is a r.v. Z on the nonnegative integers such that

$$\lambda_n = E(Z - n | Z \geqq n), \qquad n = 0, 1, 2, \cdots.$$

Consider an infinite sequence of independent coin tosses with probability $(1+\lambda_n)^{-1}$ of falling heads for the nth coin; let A_n denote the corresponding event.

THEOREM 3.4. *In order that $\{\lambda_n\}$ is a* MRL *sequence* i) *it is sufficient that $P(A_n$ i.o.$) = 1$.* ii) *if $\inf_n \lambda_n > 0$, then $\sum_n \lambda_n^{-1} = \infty$ is necessary and sufficient.*

Proof. Consider the r.v. Z such that

$$P(Z \geqq n) = \prod_{i=0}^{n-1} \frac{\lambda_i}{1+\lambda_i}, \qquad n = 1, 2, \cdots,$$
$$P(Z = 0) = \frac{1}{1+\lambda_0}. \tag{3.8}$$

Note that $P(A_n \text{ i.o.}) = 1 \Leftrightarrow \sum_n (1+\lambda_n)^{-1} = \infty$ by the Borel–Cantelli lemma; the latter implies that (3.8) is a proper tail since

$$P(Z \geqq n) = \prod_{i=0}^{n-1} [1 - (1+\lambda_i)^{-1}] < \exp\left\{-\sum_{i=0}^{n-1} (1+\lambda_i)^{-1}\right\}$$

and we can check that Z has MRL λ_n. If $\inf_n \lambda_n > 0$, set $\alpha = \inf_n \{\lambda_n/(1+\lambda_n)\}$. Then $0 < \alpha < 1$ and ii) follows by noting that

$$\exp\left(-\sum_{i=0}^{n-1} \lambda_i^{-1}\right) < \prod_{i=0}^{n-1} (1+\lambda_i^{-1})^{-1} = P(Z \geqq n)$$
$$< \exp\left\{-\sum_{i=0}^{n-1} (1+\lambda_i)^{-1}\right\} < \exp\left(-\alpha \sum_{i=0}^{n-1} \lambda_i^{-1}\right). \quad \square$$

Note that $\sum \lambda_n^{-1} = \infty$ remains necessary even when $\inf \lambda_n = 0$. Whenever λ_n is a MRL sequence, we have $\lambda_0 = EZ \in (0, \infty)$. The condition $\inf \lambda_n > 0$ is often free, e.g., if $\lambda_n \uparrow$, i.e., if Z is IMRL. Our theorem implies an apparently surprising conclusion which is useful in constructing MRL sequences: For any "pure birth process" $\{X(t): t > 0\}$ with

honest transient solutions (i.e., $P(X(t)<\infty)=1$ for all $t>0$), the transition rates λ_n, if bounded away from zero, is always a MRL sequence.

As a final illustration, consider a critical Galton–Watson process $\{Z_n: n = 0, 1, 2, \cdots\}$, with infinite mean time (T) to extinction, originating from a single ancestor ($Z_0 = 1$ almost surely). Then

$$c_n \overset{\text{def}}{=} E(Z_n | Z_n \geqq 1), \qquad n \geqq 1 \tag{3.9}$$

is a MRL sequence. This is so, since

$$1 = EZ_n = P(Z_n > 0)E(Z_n | Z_n > 0)$$

implies $P(T>n) = P(Z_n>0) = c_n^{-1}$, $n \geqq 1$, so that c_n is increasing, a forteriori $\inf_{n\geqq 1} c_n = 1/P(T>1) > 0$, and further that

$$\sum_{n=1}^{\infty} c_n^{-1} = ET - 1 = \infty$$

since ET diverges. Thus the necessary and sufficient condition in Theorem 3.4 holds.

Note that, if $g_n(j) = E(Z_n - j | Z_n \geqq j)$, $j \geqq 0$, is the MRL of Z_n, then $c_n = 1 + g_n(1)$, $n \geqq 1$, although this does not tell us that c_n in (3.9) is itself an increasing MRL sequence. Finally since $c_1 \geqq 1$, it follows from the above arguments that the modified sequence

$$c_n = \begin{cases} 1 & \text{if } n = 0, \\ E(Z_n | Z_n \geqq 1) & \text{if } n \geqq 1, \end{cases}$$

is still an IMRL sequence.

Acknowledgment. I am grateful to Samuel Kotz for his comments on an earlier draft of the manuscript.

REFERENCES

[1] K. B. ATHREYA AND P. E. NEY, *Branching Processes*, Springer-Verlag, New York, 1970.
[2] R. E. BARLOW AND F. PROSCHAN, *Statistical Theory of Reliability–Probability Models*, Holt, Rinehart and Winston, New York, 1975.
[3] D. J. BARTHOLOMEW, *Stochastic Models for Social Processes*, 2nd ed., John Wiley, New York, 1973.
[4] W. FELLER, *An Introduction to Probability Theory and Its Applications*, Vol. II, John Wiley, New York, 1966.
[5] I. A. IBRAGIMOV, *On the composition of unimodal distributions*, Theory Prob. Appl., 1 (1956), pp. 255–260.
[6] S. KARLIN, *Optimal policy for selling an asset*, in Studies in Applied Probability and Management Science, K. J. Arrow, S. Karlin and H. Scarf, eds., Stanford University Press, Stanford, CA, 1962.
[7] I. MEILIJSON, *Limiting properties of the mean residual life-time function*, Ann. Math. Statist., 43 (1972), pp. 354–357.
[8] P. S. PURI AND H. RUBIN, *A characterization based on the absolute difference of two i.i.d. random variables*, Ann. Math. Statist., 41 (1970), pp. 2113–2122.
[9] E. SENETA, *The Galton–Watson process with mean one*, J. Appl. Prob. , 4 (1967), pp. 489–495.
[10] B. H. SINGER AND S. SPILERMAN, *Social mobility models for heterogeneous populations*, in Sociological Methodology, H. L. Costner, ed., Jossey-Bass, San Francisco, 1973–74.
[11] G. B. SWARZ, *The mean residual lifetime function*, IEEE Trans. Reliability, R-22 (1973), pp. 108–109.
[12] G. S. WATSON AND W. T. WELLS, *On the possibility of improving the mean useful life of items by eliminating those with short lives*, Technometrics, 3 (1961), pp. 281–298.
[13] G. H. WEISS AND M. DISHON, *Some economic problems related to burn-in programs*, IEEE Trans. Reliability, R-20 (1971), pp. 190–195.
[14] S. KOTZ AND D. N. SHANBHAG, *Some new approaches to probability distributions*, Adv. Appl. Prob., 12 (1980), pp. 903–921.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 1, March 1982

0196-5212/82/0301-0007 $01.00/0

ON PACKING TWO-DIMENSIONAL BINS*

F. R. K. CHUNG†, M. R. GAREY† AND D. S. JOHNSON†‡

Abstract. Suppose we are given a set L of rectangular items and wish to pack them into identical rectangular bins, so that no two items overlap and so that the number of bins used is minimized. This generalization of the standard one-dimensional bin packing problem models problems arising in a variety of applications, from truck loading to the design of VLSI chips. We propose a hybrid algorithm, based on algorithms for simpler bin packing problems, and show that proof techniques developed for the simpler cases can be combined to prove close bounds on the worst case behavior of the new hybrid. These are the first such close bounds obtained for this problem.

1. Two-dimensional bin packing. Let $L = \{r_1, r_2, \cdots, r_n\}$ be a set of rectangles, each rectangle r having height $h(r)$ and width $w(r)$. A *packing* P of L into a collection $\{B_1, B_2, \cdots, B_m\}$ of $H \times W$ rectangular bins is an assignment of each rectangle to a bin and a position within that bin such that (a) each rectangle is contained entirely within its bin, with its sides parallel to the sides of the bin, and (b) no two rectangles in a bin overlap. See Fig. 1 for an example of such a packing. In this paper we also

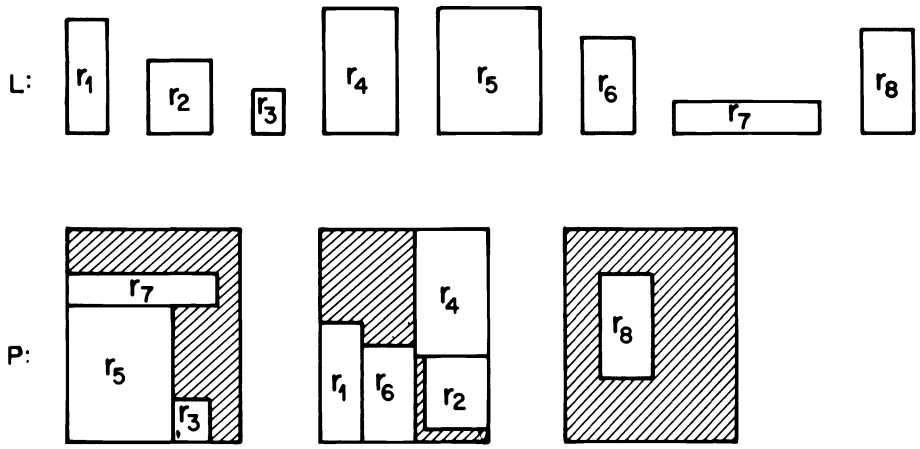


FIG. 1. *Example of a packing P of a list L of rectangles into* 3 *bins with* $H = 20$, $W = 16$. *Rectangle dimensions are* 11×4, 7×6, 4×3, 12×7, 13×10, 9×5, 3×14, *and* 10×5.

assume that the orientations of the rectangles cannot be changed: the width of a rectangle must be aligned with the width of the bin. (The case when 90° rotations are allowed will be discussed in the conclusion.) In what follows, we shall assume that the bin dimensions H and W have been fixed and hence all packings are into bins of that size.

If P is a packing, let $|P|$ denote the number of nonempty bins in P. Given a list L, let OPT (L) be defined to be $\min\{|P| : P \text{ is a packing of } L\}$. We are interested in finding packings P with $|P|$ close to OPT (L). (Determining OPT (L), given L, is an NP-hard problem [1, Ch. 10], [7], and so it is unlikely that we can find *optimal* packings efficiently.)

* Received by the editors May 6, 1981.

† Bell Laboratories, Murray Hill, New Jersey 07974.

‡ The work of this author was supported in part by the Computer Sciences Department, University of Wisconsin, Madison, Wisconsin 53706.

This problem is related to two simpler and well-studied packing problems: one-dimensional bin packing [9], [10] and two-dimensional strip packing [2], [3], [4]. The first is equivalent to the special case of our problem in which $w(r) = W$ for all $r \in L$. In the second we are once more given an arbitrary set of rectangles, but this time we are asked to pack them into a strip of width W so as to minimize the height of the strip used. Although considerable progress has been made in analyzing the worst case behavior of algorithms for these two simpler problems, until now there has been little success in extending the results to the case of two-dimensional bin packing. In this paper we make a start in this direction by proposing an appealing hybrid algorithm and obtaining close bounds on its asymptotic worst case behavior.

2. Asymptotic worst case analysis. We measure the asymptotic worst case behavior of an algorithm A by the quantity R_A^∞, defined as follows: Let $A(L)$ be the value of the packing obtained by applying A to L. ($A(L)$ would be either the number of bins or the strip height, depending on the problem.) Let OPT (L) be the corresponding optimal value. We then define $R_A(L) \equiv A(L)/\text{OPT}(L)$, $R_A^n \equiv \max\{R_A(L)\colon L$ satisfies OPT $(L) = n\}$, and finally $R_A^\infty = \limsup_{n\to\infty} R_A^n$. The closer R_A^∞ is to one, the better is the asymptotic worst case behavior of A.

Our hybrid algorithm is built from algorithms already developed for the simpler cases. The FIRST FIT algorithm (FF) for the one-dimensional problem places the first item at the bottom of the first bin, and thereafter places each item in turn in the lowest indexed bin which has room for it. In [9], [10] it is shown that $R_{FF}^\infty = \frac{17}{10}$. The FIRST FIT DECREASING algorithm (FFD) is the same as FF, except that the items to be packed are initially reordered so that $h(r_1) \geqq h(r_2) \geqq \cdots \geqq h(r_n)$. For this algorithm we have [9], [10] that $R_{FFD}^\infty = \frac{11}{9} = 1.222\cdots$.

We shall be using FFD together with a strip packing algorithm based on FF, which we call FIRST FIT BY DECREASING HEIGHT (FFDH). The FFDH algorithm constructs a packing in which the strip is stratified into *blocks*, each block running the full width of the strip and resting on the top of the previous block (the first block rests on the bottom of the strip). Within the blocks, rectangles are packed linearly, each with its bottom edge resting on the bottom of the block. The height of a block is the height of the tallest rectangle it contains. Algorithm FFDH works by first reordering the set L of rectangles so that $h(r_1) \geqq h(r_2) \geqq \cdots \geqq h(r_n)$ and then proceeding as follows: Place the first rectangle left-justified in the first block. Thereafter the rectangles are assigned in turn, each rectangle being placed as far to the left as possible in the lowest block which has room for it along its bottom edge. A new block is started on top of the current top block whenever the rectangle will not fit in any of the current blocks. See Fig. 2 for the FFDH packing of the rectangles of Fig. 1, appropriately reindexed by height.

Note that if all the rectangles were the same height, FFDH would be equivalent to FF, with the blocks playing the role of bins. In [4] it is shown that the fact that rectangles may have differing heights is not as damaging as one might think, for $R_{\text{FFDH}}^\infty = R_{\text{FF}}^\infty = \frac{17}{10}$.

3. A hybrid algorithm. Our hybrid algorithm is now quite easily described. First create a strip packing for L using FFDH and strip width W, thereby obtaining a collection $\{b_1, b_2, \cdots, b_k\}$ of blocks of nonincreasing heights $h_1 \geqq h_2 \geqq \cdots \geqq h_k$, each containing a subset of the rectangles. If we view these blocks as a new collection of rectangles $L' = \{b_1, b_2, \cdots, b_k\}$ with $h(b_i) = h_i$ and $w(b_i) = W$, $1 \leqq i \leqq k$, we have an instance of the one-dimensional problem and can apply FFD to pack the blocks (and hence the rectangles they contain) into $H \times W$ bins. See Fig. 3, where FFD has been

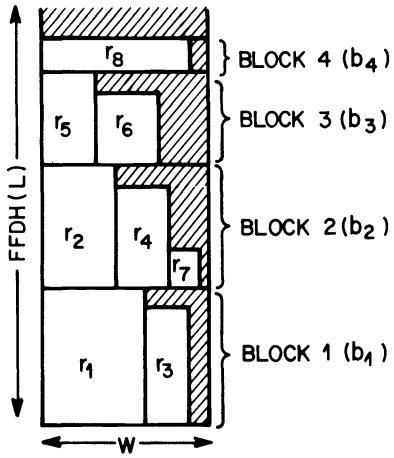


FIG. 2. *Example of an* FFDH *packing of a list L of rectangles with dimensions* 13×10, 12×7, 11×4, 10×5, 9×5, 7×6, 4×3, *and* 3×14.

applied to the blocks of the strip packing in Fig. 2. We call this hybrid algorithm HYBRID FIRST FIT (HFF). Our main result is

$$2.022 \leqq R^{\infty}_{\mathrm{HFF}} \leqq 2.125.$$

In Fig. 4 we present a schematic for instances L of two-dimensional bin packing with arbitrarily large values of OPT (L) for which HFF $(L) = \frac{91}{45}(\mathrm{OPT}\,(L)-1)$. These instances will thus imply the lower bound $R^{\infty}_{\mathrm{HFF}} \geqq 2.0222\cdots$. The optimal packing is shown in Fig. 4(a) and consists of three types of bins: $42n$ bins containing items of types A, B, and E, packed as illustrated, followed by $48n$ bins containing items of types A, C, D, and E, packed as illustrated, followed by a single bin containing a single item of type A. The precise dimensions of the items are as follows (δ and ε to be specified later):

$$A\text{-item in bin } j\text{: height} = \tfrac{1}{2}+\varepsilon,\ \text{width} = \begin{cases} \frac{1}{6}+4^j\delta & \text{if } j \text{ odd},\\ \frac{1}{6}-4^j\delta & \text{if } j \text{ even};\end{cases}$$

$$B\text{-item in bin } j\text{: height} = \tfrac{1}{4}+2\varepsilon,\ \text{width} = \begin{cases} \frac{1}{3}-(4^j+1)\delta & \text{if } j \text{ odd},\\ \frac{1}{3}+(4^j-1)\delta & \text{if } j \text{ even};\end{cases}$$

$$C\text{-item in bin } j\text{: height} = \tfrac{1}{2}+\varepsilon,\ \text{width} = \begin{cases} \frac{1}{3}-(4^j+1)\delta & \text{if } j \text{ odd},\\ \frac{1}{3}+(4^j-1)\delta & \text{if } j \text{ even};\end{cases}$$

All D-items have height $= \frac{1}{4}+\varepsilon$, width $= \frac{1}{2}+\delta$;

All E-items have height $= \frac{1}{4}-2\varepsilon$, width $= \frac{1}{2}+\delta$.

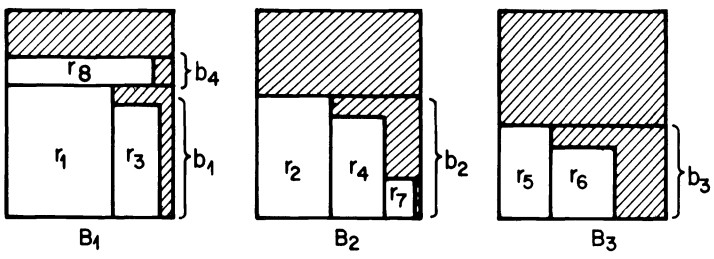
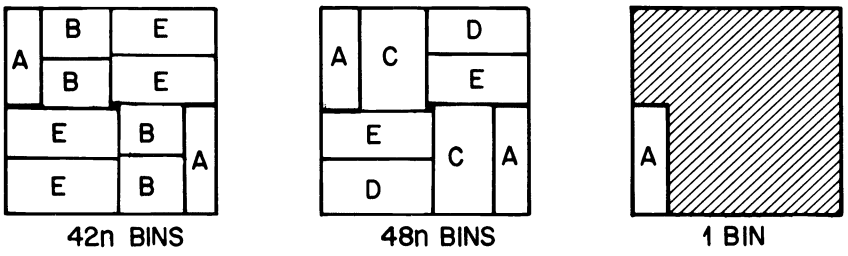
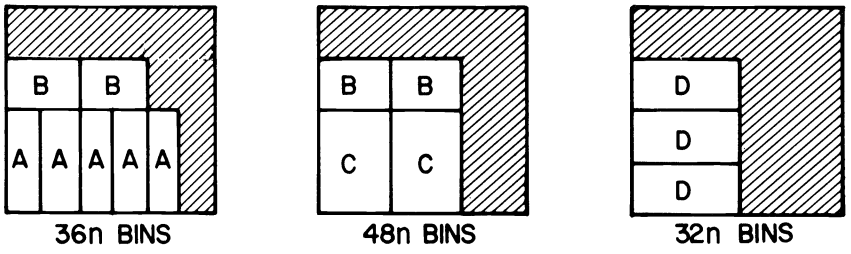
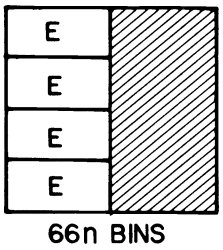


FIG. 3. HFF *packing based on the* FFDH *packing of Fig.* 2.



(a) OPTIMAL PACKING: OPT(L)=90n+1.





(b) HFF PACKING: HFF(L)=182n.

FIG. 4. *Schematic of packings of lists L with* HFF $(L)=\frac{91}{45}(\mathrm{OPT}\,(L)-1)$.

The reader may readily verify that if we choose ε and δ so that $0<\varepsilon<\frac{1}{16}$ and $0<\delta<4^{-50n}$, the items can be packed as claimed.

For the application of HFF, these items must be ordered by decreasing height. We assume that ties among items of the same height are broken so that the items are ordered as follows: First come the A-items, in reverse order, with the first A-item from bin $2i+1$ replaced by the first from bin $2i+3$, $0\leqq i\leqq 45n-1$. To illustrate this, here is a list of the values for the first 20 A-items of $w(r)-\frac{1}{6}$ (we let $m=90n$):

$$-4^{m}\delta,\ -4^{m}\delta,\ +4^{m+1}\delta,\ +4^{m-1}\delta,\ -4^{m-2}\delta,$$
$$-4^{m-2}\delta,\ +4^{m-1}\delta,\ +4^{m-3}\delta,\ -4^{m-4}\delta,\ -4^{m-4}\delta,$$
$$+4^{m-3}\delta,\ +4^{m-5}\delta,\ -4^{m-6}\delta,\ -4^{m-6}\delta,\ +4^{m-5}\delta,$$
$$+4^{m-7}\delta,\ -4^{m-8}\delta,\ -4^{m-8}\delta,\ +4^{m-7}\delta,\ +4^{m-9}\delta.$$

Note that after each set of five items FFDH would start a new block: The sum of the first five exceeds $\frac{5}{6}+4^{m}\delta$ and hence none of the remaining items will fit in the gap, and similar arguments hold for all remaining sets of five. Thus, since there are a total

of $180n+1$ type A-items, FFDH will create $36n$ "A-blocks" of 5 A-items each (the last A-item, having width $\frac{1}{6}+4\delta$, will be postponed until after the C-items, and can be ignored since it will just fall in the first C-block).

The C-items follow the A-items, and are ordered so that they will go two per block. The values of $w(r)-\frac{1}{3}$ for the first eight are

$$+(4^m-1)\delta,\ -(4^{m-1}+1)\delta,\ +(4^m-1)\delta,\ -(4^{m-1}+1)\delta,$$
$$+(4^{m-2}-1)\delta,\ -(4^{m-3}+1)\delta,\ +(4^{m-2}-1)\delta,\ -(4^{m-3}+1)\delta.$$

The reader should be able to see that this type of ordering will yield $48n$ blocks of 2 C-items each out of the total of $96n$ C-items. Similar tricks are played with the $168n$ B-items which follow next, yielding $84n$ blocks of two B-items each. (Note that sizes are arranged so that no B-item is narrow enough to fit in a block of C-items.)

Finally, the list concludes with the $96n$ D-items, each going in a block by itself, followed by the $264n$ E-items, each going in a block by itself.

The reader may now verify that when HFF applies FFD to the blocks thus created, the packing of Fig. 4(b) will result, using $182n$ bins or $\frac{91}{45}(\mathrm{OPT}\,(L)-1)$ as claimed. Note also that the bad behavior illustrated here is not dependent on our ability to order items of the same height in the worst possible way, since by appropriately shaving the height of the items we can insure that the given order is *forced* by the decreasing height rule, without changing the natures of the optimal and HFF packings.

The upper bound on $R^{\infty}_{\mathrm{HFF}}$ comes from the following theorem:

THEOREM 1. *For any list L of rectangles,* $\mathrm{HFF}\,(L)<\frac{17}{8}\,\mathrm{OPT}\,(L)+5$.

Proof. Suppose that L is a counter-example. By normalizing widths and heights, we may assume without loss of generality that $W=H=1$ and $0\leqq w(r),\ h(r)\leqq 1$ for all $r\in L$. Let us further assume that L is a counter-example containing the minimum possible number of rectangles.

We rely on three results about the one-dimensional bin packing and two-dimensional strip packing problems. Let $f: L\to[0,\frac{8}{5}]$ be a weighting function defined as follows:

$$f(r)=\begin{cases}(\frac{6}{5})\cdot w(r) & \text{if } 0\leqq w(r)\leqq\frac{1}{6},\\ (\frac{9}{5})\cdot w(r)-\frac{1}{10} & \text{if } \frac{1}{6}<w(r)\leqq\frac{1}{3},\\ (\frac{6}{5})\cdot w(r)+\frac{1}{10} & \text{if } \frac{1}{3}<w(r)\leqq\frac{1}{2},\\ (\frac{6}{5})\cdot w(r)+\frac{4}{10} & \text{if } \frac{1}{2}<w(r)\leqq 1.\end{cases}$$

LEMMA 1 [6]. *If $R\subseteq L$ and $w(R)\equiv\sum_{r\in R}w(r)\leqq 1$, then $f(R)\equiv\sum_{r\in R}f(r)\leqq\frac{17}{10}$.*

LEMMA 2 [6]. *Suppose $R\subseteq L$ and $\{R_1, R_2, \cdots, R_m\}$ is a partition of R into disjoint nonempty sets such that for all integers i and j with $1\leqq i<j\leqq m$, $r\in R_j$ implies $w(r)>H1-w(R_i)$. Then $f(R)\geqq m-1$.*

LEMMA 3 [4]. *Suppose $\mathrm{OPT}_S\,(L)$ is the minimum possible strip height H' such that L can be packed into a strip of width 1 and height H'. Then* $\mathrm{FFDH}\,(L)\leqq\frac{17}{10}\,\mathrm{OPT}_S\,(L)+1$.

Given $L=\{r_1, r_2, \cdots, r_n\}$, we now show that $\mathrm{HFF}\,(L)<\frac{17}{8}\,\mathrm{OPT}\,(L)+5$, in contradiction of our assumption that L was a counter-example. Let P_{HFF} be the HFF packing of L and P_{OPT} be an optimal packing. Let x denote the height of the tallest block in the last bin of P_{HFF}. Since L is a minimum counter-example, we may assume that all rectangles $r_i\in L$ have height at least x: The number and heights of blocks of height x or greater would not be affected by deleting all rectangles shorter than x, so that the number of bins required by HFF would not decrease, whereas the number of bins

required by an optimal packing could not increase. Thus if L contained any rectangle shorter than x, a counter-example with fewer items would exist, contradicting the minimality of L.

Our proof divides into four cases, depending on the value of x. We shall treat the cases in order of difficulty.

Case 1. $x \leqq \frac{1}{5}$. In this case all but the last bin of P_{HFF} must contain blocks whose total height is at least $\frac{4}{5}$. Thus, by Lemma 3,

$$\tfrac{4}{5}(\mathrm{HFF}(L)-1 \leqq \mathrm{FFDH}(L) \leqq \tfrac{17}{10}\,\mathrm{OPT}_S(L)+1 \leqq \tfrac{17}{10}\,\mathrm{OPT}(L)+1,$$

where the last inequality results from the fact that one way to pack a strip of width 1 with L is to pack L into $\mathrm{OPT}(L)$ bins of width and height 1 and then pile them one on top of another. From this we conclude that

$$\mathrm{HFF}(L) \leqq \tfrac{5}{4} \cdot \tfrac{17}{10}\,\mathrm{OPT}(L)+\tfrac{9}{4} < \tfrac{17}{8}\,\mathrm{OPT}(L)+5,$$

as desired.

In the remaining cases we assume that $x > \frac{1}{5}$ and so can divide the items of L into the following classes:

$$X_1 = \{r_i : h(r_i) > 1-x\},$$

$$X_2 = \left\{r_i : 1-x \geqq h(r_i) > \frac{1}{2}\right\},$$

$$X_3 = \left\{r_i : \frac{1}{2} \geqq h(r_i) > \frac{1-x}{2}\right\},$$

$$X_4 = \left\{r_i : \frac{1-x}{2} \geqq h(r_i) > \frac{1}{4}\right\},$$

$$X_5 = \left\{r_i : \frac{1}{4} \geqq h(r_i) \geqq x\right\}.$$

We shall say a block is of "type X_i" if its tallest item is from X_i.

Let $B_1, B_2, \cdots, B_l$ denote the bins of P_{HFF} in order, where $l = \mathrm{HFF}(L)$. For $1 \leqq i \leqq 5$, let β_i denote the set of bins whose tallest block $_i$, and let $N_i = |\beta_i|$. Note that all bins from β_i precede all bins from β_{i+1}, $1 \leqq i \leqq 4$.

Case 2. $x > \frac{1}{3}$. If $x > \frac{1}{3}$ then $(1-x)/2 < \frac{1}{3}$ and so $N_4 = N_5 = 0$. Let us look at an arbitrary bin B in P_{OPT} and imagine lines drawn through it at heights $\frac{1}{3}$ and $\frac{2}{3}$. Let $S_1(B)$ be the set of items from X_1 in B. Let $S_{23}(B, 1)$ be the set of items from X_2 and X_3 in B whose interiors are traversed by the line at height $\frac{1}{3}$, and let $S_{23}(B, 2)$ be the set of items from X_2 and X_3 in B whose interiors are traversed by the line at height $\frac{2}{3}$ but not by the line at $\frac{1}{3}$. Note that since all items in L are of height exceeding $\frac{1}{3}$, every item in B must be in precisely one of these three sets.

Now observe that, since every item in X_1 has height exceeding $1-x$, no vertical line through B can traverse the interiors of both an item from $S_1(B)$ and one from $S_{23}(B, 1) \cup S_{23}(B, 2)$. We thus have

$$w(S_1(B)) + w(S_{23}(B, 1)) \leqq 1, \tag{2.1}$$

$$w(S_1(B)) + w(S_{23}(B, 2)) \leqq 1. \tag{2.2}$$

Using Lemma 1 and summing over all bins B of P_{OPT} we conclude that

$$2f(X_1) + f(X_2) + f(X_3) \leqq 2 \cdot \tfrac{17}{10}\,\mathrm{OPT}(L) \leqq \tfrac{17}{5}\,\mathrm{OPT}(L). \tag{2.3}$$

We now turn to the HFF packing. The bins of β_1 each contain one block, that block having height exceeding $1-x$, and these blocks induce a partition on X_1 which obeys the hypotheses of Lemma 2. Thus

$$f(X_1) \geqq N_1 - 1. \tag{2.4}$$

None of the remaining bins contains a block of type X_1 and so the fact that a block of height x went into the last bin means that all except that last bin must contain at least (and hence exactly) two blocks. Letting X'_{23} denote the subset of $X_2 \cup X_3$ which is contained in these bins, and ordering the blocks in the same order as they were created by FFDH, we see that these $2(N_2+N_3)-1$ blocks induce a partition of X'_{23} which obeys the hypotheses of Lemma 2. Therefore

$$f(X_2)+f(X_3) \geqq f(X'_{23}) \geqq 2(N_2+N_3)-2. \tag{2.5}$$

Substituting (2.4) and (2.5) into (2.3) we obtain

$$2(N_1-1)+2(N_2+N_3-1) \leqq \tfrac{17}{5}\,\mathrm{OPT}\,(L)$$

or

$$\mathrm{HFF}\,(L) = N_1+N_2+N_3 \leqq \tfrac{17}{10}\,\mathrm{OPT}\,(L)+2 < \tfrac{17}{8}\,\mathrm{OPT}\,(L)+5,$$

as desired.

Case 3. $\frac{1}{4} < x \leqq \frac{1}{3}$. In this case $N_5 = 0$. Let us once again consider a bin B in the optimal packing. This time we imagine 7 horizontal lines drawn through B: two (identical) lines at height x, one at height $(1-x)/2$, one at $\frac{1}{2}$, one at $(1+x)/2$, and two (identical) lines at $1-x$. It is easy to verify that, given these lines, each rectangle from X_1 in B will have its interior traversed by all 7 lines. Similarly, rectangles from X_2, X_3 and X_4 will have their interiors traversed by at least 4, 3, and 2 lines, respectively. Let $S_i(B, j)$ be the set of rectangles from X_i whose interiors are traversed by the jth line, $1 \leqq i \leqq 4$, $1 \leqq j \leqq 7$. We then have, for each j, $1 \leqq j \leqq 7$,

$$\sum_{i=1}^{4} w(S_i(B, j)) \leqq 1.$$

Lemma 1 then yields for each j, $1 \leqq j \leqq 7$,

$$\sum_{i=1}^{4} f(S_i(B, j)) \leqq \tfrac{17}{10}.$$

Summing over all bins B of P_{OPT} we thus conclude

$$7f(X_1)+4f(X_2)+3f(X_3)+2f(X_4) \leqq 7 \cdot \tfrac{17}{10}\,\mathrm{OPT}\,(L). \tag{3.1}$$

Turning to the HFF packing, let $\beta_{2,3}$ be the set of bins from β_2 that, in addition to containing a block of type X_2, also contain a block of type X_3. Since a block of type X_2 has height at most $1-x$ and since the block of height x in the last bin did not fit in any earlier bin, every bin in $\beta_{2,4} = \beta_2 - \beta_{2,3}$ must contain a block of type X_4. Let $N_{2,i} = |\beta_{2,i}|$ for $i \in \{3, 4\}$.

Applying Lemma 2 to the partitions of X_1 and X_2 induced by the bins of β_1 and β_2 respectively, we obtain

$$f(X_1) \geqq N_1 - 1, \tag{3.2}$$

$$f(X_2) \geqq N_2 - 1. \tag{3.3}$$

There are at least $N_{2,3}+2N_3-1$ blocks of type X_3: one in each bin of $\beta_{2,3}$ and two in all but possibly the last bin of β_3. If we let X_3' be the subset of X_3 contained in these blocks and apply Lemma 2 to the partition of X_3' induced by these blocks, we obtain

$$f(X_3) \geqq f(X_3') \geqq N_{2,3}+2N_3-2. \tag{3.4}$$

Finally, consider the blocks of type X_4. There are at least $N_{2,4}+3N_4-2$ of these: one in each bin of $\beta_{2,4}$ and three in each bin of β_4 except the last. (A nonfinal bin from β_4 cannot have height less than $1-x$, and since no block of type X_4 has height exceeding $(1-x)/2$, each such bin must contain at least three blocks.) Lemma 2 thus yields

$$f(X_4) \geqq N_{2,4}+3N_4-3. \tag{3.5}$$

Substituting (3.2) through (3.5) in (3.1) yields

$$7N_1+6N_2+6N_3+6N_4-23 \leqq 7 \cdot \tfrac{17}{10}\,\mathrm{OPT}\,(L)$$

or

$$\mathrm{HFF}\,(L) = N_1+N_2+N_3+N_4 \leqq \tfrac{7}{6} \cdot \tfrac{17}{10}\,\mathrm{OPT}\,(L)+\tfrac{23}{6} < \tfrac{17}{8}\,\mathrm{OPT}\,(L)+5,$$

as desired.

Case 4. $\frac{1}{5} < x \leqq \frac{1}{4}$. We divide this case into two subcases, depending on the value of N_4.

Subcase 4.1. $N_4 = 0$. The total height of all blocks in P_{HFF} is bounded by $(1-x)(N_1+N_2+N_3)+4x(N_5-1)$ since all bins except the last must have total block height at least $1-x$, and all bins of β_5 except the last must contain four blocks. By Lemma 3 we thus have

$$(1-x)(N_1+N_2+N_3)+4x(N_5-1) \leqq \tfrac{17}{10}\,\mathrm{OPT}\,(L)+1. \tag{4.1}$$

Furthermore, by the argument used in Case 2, we have

$$N_1+N_2+N_3 \leqq \tfrac{17}{10}\,\mathrm{OPT}\,(L)+2. \tag{4.2}$$

Using (4.1) and (4.2) we then can derive the following:

$$\begin{aligned} 4xl &= 4x(N_1+N_2+N_3+N_5) \\ &\leqq 4x(N_1+N_2+N_3)+\tfrac{17}{10}\,\mathrm{OPT}\,(L)+1-(1-x)(N_1+N_2+N_3)+4x \\ &\leqq (5x-1)(\tfrac{17}{10}\,\mathrm{OPT}\,(L)+2)+\tfrac{17}{10}\,\mathrm{OPT}\,(L)+1+4x \\ &\leqq (5x)\tfrac{17}{10}\,\mathrm{OPT}\,(L)+14x-1 \end{aligned}$$

and hence $l = \mathrm{HFF}\,(L) \leqq \frac{5}{4} \cdot \frac{17}{10}\,\mathrm{OPT}\,(L)+\frac{7}{2} < \frac{17}{8}\,\mathrm{OPT}\,(L)+5$, as desired.

Subcase 4.2. $N_4 > 0$. Consider a bin B in P_{OPT} and this time imagine seven horizontal lines drawn through it, at heights $j/8$, $1 \leqq j \leqq 7$. Then rectangles from classes X_1, X_2, X_3, X_4, and X_5 have their interiors traversed by at least 6, 4, 3, 2, and 1 lines respectively, since $1-x \geqq \frac{3}{4}$ and $(1-x)/2 \geqq \frac{3}{8}$.

Letting $S_i(B, j)$ be the set of rectangles from X_i whose interiors are traversed by the jth line, $1 \leqq i \leqq 5$, $1 \leqq j \leqq 7$, we then have for each j, $1 \leqq j \leqq 7$,

$$\sum_{i=1}^{5} w(S_i(B, j)) \leqq 1.$$

Lemma 1 thus yields $\sum_{i=1}^{5} f(S_i(B, j)) \leqq \frac{17}{10}$, and summing over all bins B of P_{OPT} we obtain

$$6f(X_1)+4f(X_2)+3f(X_3)+2f(X_4)+f(X_5) \leqq 7\tfrac{17}{10}\,\mathrm{OPT}\,(L). \tag{4.3}$$

We now turn to the HFF packing. Since $N_4 > 0$, there is a block of type X_4 which did not fit in any bin from β_2 or any bin from β_4 except the last. Thus any bin from class β_2 or any bin (except the last) from class β_4 that contains a block of type X_5 must contain blocks whose total height is at least $1-(1-x)/2+x = (1+3x)/2$. Let us partition the bins in β_2 and β_4 as follows:

Any bin in β_2 must contain at least one block in addition to its block of type X_2. Let $\beta_{2,j}$, $3 \leqq j \leqq 5$, be the subset of bins from β_2 whose second block is of type X_j (there may be a third block, but we ignore it in forming the partition). Similarly, any bin in β_4 must contain at least three blocks. Let $\beta_{4,5}$ be the set of bins in β_4, other than the last, for which the third block is of type X_5, and let $\beta_{4,4} = \beta_4 - \beta_{4,5}$. Letting $N_{i,j} = |\beta_{i,j}|$, we then have

$$\begin{aligned}(1-x)(N_1+N_{2,3}+N_{2,4}+N_3+N_{4,4})+\left(\frac{1+3x}{2}\right)(N_{2,5}+N_{4,5})+4x(N_5-1)\\ \leqq \mathrm{FFDH}\,(L) \leqq \tfrac{17}{10}\,\mathrm{OPT}\,(L)+1.\end{aligned} \tag{4.4}$$

Our next inequalities are obtained by applying Lemma 1 to the blocks of type X_i, $1 \leqq i \leqq 5$, as in previous cases, using the facts that all but the last bin in β_3 contain 2 blocks of type X_3, all but the last bin in β_4 contain either 3 blocks of type X_4 (if in $\beta_{4,4}$) or two (if in $\beta_{4,5}$), and all but the last bin in β_5 contain 4 blocks of type X_5:

$$f(X_1) \geqq N_1 - 1, \tag{4.5}$$

$$f(X_2) \geqq N_2 - 1, \tag{4.6}$$

$$f(X_3) \geqq N_{2,3} + 2N_3 - 2, \tag{4.7}$$

$$f(X_4) \geqq N_{2,4} + 3N_{4,4} + 2N_{4,5} - 3, \tag{4.8}$$

$$f(X_5) \geqq N_{2,5} + N_{4,5} + 4N_5 - 4. \tag{4.9}$$

Now a final dose of symbol manipulation yields the desired result. Combining (4.3) and (4.5) through (4.9) we obtain

$$\begin{aligned}6(N_1)+6(N_{2,3}+N_{2,4})+5N_{2,5}+6N_3+6N_{4,4}+5N_{4,5}+4N_5\\ \leqq 7 \cdot \tfrac{17}{10}\,\mathrm{OPT}\,(L)+26.\end{aligned} \tag{4.10}$$

Multiplying (4.4) by 2 and (4.10) by $(5x-1)$ and then adding we obtain

$$\begin{aligned}(2(1-x)+6(5x-1))(N_1+N_{2,3}+N_{2,4}+N_3+N_{4,4})\\ +((1+3x)+5(5x-1))(N_{2,5}+N_{4,5})+(8x+4(5x-1))(N_5)\\ \leqq \tfrac{17}{10}\,\mathrm{OPT}\,(L)(7(5x-1)+2)+26(5x-1)+8x+2,\end{aligned}$$

that is,

$$\begin{aligned}(28x-4)(N_1+N_{2,3}+N_{2,4}+N_{2,5}+N_3+N_{4,4}+N_{4,5}+N_5)\\ \leqq \tfrac{17}{10}\,\mathrm{OPT}\,(L)(35x-5)+138x-24\end{aligned}$$

or

$$\begin{aligned}\mathrm{HFF}(L) &< \tfrac{5}{4} \cdot \tfrac{17}{10}\,\mathrm{OPT}(L) + 5 \\ &= \tfrac{17}{8}\,\mathrm{OPT}(L) + 5,\end{aligned}$$

as desired.

Thus in all cases $\mathrm{HFF}(L) < \frac{17}{8}\,\mathrm{OPT}(L) + 5$, in contradiction to our assumption that a counter-example exists. The theorem has been proved. □

4. Directions for further research. By showing that close bounds can be obtained on the asymptotic worst case behavior of two-dimensional bin packing algorithms, we hope to encourage researchers to design other algorithms and investigate their behavior. Algorithms based on the "bottom-left" strip packing rule introduced in [3] are particularly attractive candidates for analysis. Although the bottom-left algorithms are all asymptotically worse than FFDH in the strip packing environment, they may well be more competitive for two-dimensional bin packing. There is also the possibility of constructing better hybrid algorithms. FFDH is not the best heuristic known for strip packing. An algorithm is presented in [2] with $R_A^\infty \leqq \frac{5}{4}$ (although the structure of its packings is much more complicated than that for FFDH). Similarly, FFD has recently been improved on in the one-dimensional case by a modified algorithm [8] with $R_A^\infty = 1.18333\cdots$.

A second line of attack would be to design and analyze algorithms which could make use of the fact that, in some applications, 90° rotations of rectangles might be allowable. Algorithm HFF would still be applicable in such situations, assuming all rectangles were presented in such a way that they would fit in a bin without rotation. However, the performance guarantee of Theorem 1 would not necessarily hold. Algorithms which consider the possibility of rotations might well yield improvements. Can one prove worst case bounds that reflect these improvements?

Finally, there is of course the problem of further narrowing the gap between upper and lower bounds on R_{HFF}^∞. We suspect that the upper bound can be lowered further, although we fear that a considerable blow-up in proof length might be necessary. As to the actual value of R_{HFF}^∞, we hesitate to conjecture. It is amusing to note that one possibility still left open by our bounds is $(\frac{17}{10})(\frac{11}{9}) = 2.07777\cdots$, the product of the values of R_A^∞ for the two algorithms whose combination yields the algorithm HFF, although we suspect that the actual value may be somewhat less than this.

REFERENCES

[1] A. V. AHO, J. E. HOPCROFT AND J. D. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[2] B. S. BAKER, D. J. BROWN AND H. P. KATSEFF, *A 5/4 algorithm for two-dimensional packing*, to appear.

[3] B. S. BAKER, E. G. COFFMAN, JR. AND R. L. RIVEST, *Orthogonal packings in two dimensions*, SIAM J. Comput., 9 (1980), pp. 846–855.

[4] E. G. COFFMAN, JR., M. R. GAREY, D. S. JOHNSON AND R. E. TARJAN, *Performance bounds for level-oriented two-dimensional packing algorithms*, SIAM J. Comput, 9 (1980), pp. 808–826.

[5] M. R. GAREY, R. L. GRAHAM AND D. S. JOHNSON, *On a number-theoretic bin packing conjecture*, in Proc. 5th Hungarian Combinatorics Colloquium, North-Holland, Amsterdam, 1978, pp. 377–392.

[6] M. R. GAREY, R. L. GRAHAM, D. S. JOHNSON AND A. C. YAO, *Resource constrained scheduling as generalized bin packing*, J. Combin. Theory Ser. A, 21 (1976), pp. 257–298.

[7] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability: A Guide to the Theory of* NP-*Completeness*, W. H. Freeman, San Francisco, 1979.

[8] ———, in preparation.

[9] D. S. JOHNSON, *Near-optimal bin packing algorithms*, Tech. Rep. MAC TR-109, Project MAC, Massachusetts Institute of Technology, Cambridge, MA, 1973.

[10] D. S. JOHNSON, A. DEMERS, J. D. ULLMAN, M. R. GAREY AND R. L. GRAHAM, *Worst-case performance bounds for simple one-dimensional packing algorithms*, SIAM J. Comput., 3 (1974), pp. 299–325.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 1, March 1982

0196-5212/82/0301-0008 $01.00/0

A PHASE-TYPE SEMI-MARKOV POINT PROCESS*

GUY LATOUCHE†

Abstract. A semi-Markov point process is defined for which the intervals of time between successive events have phase-type distribution. The distribution of the number of events in an interval is examined, and it is shown how the expected number of events in an interval may be efficiently computed. A stationary version of the process is analyzed. In particular, a necessary and sufficient condition, and simple sufficient conditions under which the new process is a renewal process, are determined.

Introduction. In the present paper, we define and discuss a special semi-Markov point process. It is assumed that there are N different types of intervals. The types of successive intervals are determined by a Markov chain with transition probability matrix P. If a given interval is of type i, then the following interval is of type j with probability P_{ij}. There is a substantial literature on semi-Markov point process (Pyke [10], [11], Cinlar [1], [2]). The new feature here is that the random intervals are assumed to have phase-type distributions (defined in the next section).

This work on phase-type semi-Markov point processes was motivated by our interest in queueing problems. Most of the present literature on queueing theory deals with systems for which arrivals occur according to a renewal process. From the available results, it is clear that the analysis of a queueing system with a general nonrenewal arrival process is very difficult. On the other hand, phase-type distributions have great versatility and the structure of our process is very simple. It should, therefore, be a useful tool in modeling queueing systems with nonindependent arrivals, and provide analytically or algorithmically tractable results, as in Latouche [6], and Neuts and Chakravarthy [9], where special cases are considered.

We shall give a formal definition of the process in the next section. We examine in § 2 the number of events in an interval and show how the expected number of events may be efficiently computed. The point process under consideration here is a special case of the "versatile Markovian point process" defined in Neuts [8]. Because of the special structure, we are able to present more precise results. In § 3, we analyze the correlation structure of a stationary version of the process. We determine a necessary and sufficient condition under which that process is a renewal process. Similar problems are examined in Simon [12], for Markov-renewal process with general distributions. We comment about the results in [12] at the end of § 3. In § 4, some examples such as the interrupted Poisson process are considered in further detail.

Notational convention. All vectors are represented by boldface letters. The context indicates whether they are row or column vectors. In order to facilitate the reading of the formulas, the expression $\mathbf{vw}$ always represents the inner product of a row vector $\mathbf{v}$ by a column vector $\mathbf{w}$; the expression $\mathbf{w} \cdot \mathbf{v}$ always represents the product of a column vector $\mathbf{w}$ by a row vector $\mathbf{v}$, yielding a matrix whose (i, j)th element is $w_i v_j$.

1. The phase-type semi-Markov point process.

1.1 Phase-type distributions. Phase-type distributions have been introduced by Neuts [7]. Consider an $(n+1)$-state continuous-parameter Markov process, with n

* Received by the editors July 11, 1980, and in final form May 20, 1981. This research was supported in part by the National Science Foundation under grant ENG-7908351 and by the Air Force Office of Scientific Research under grant AFOSR-77-3236.

† Laboratoire d'Informatique Théorique, Faculté des Sciences, Université Libre de Bruxelles, Boulevard du Triomphe, B-1050 Bruxelles, Belgium.

transient states and one absorbing state. Its infinitesimal generator Q is of the form

$$Q = \begin{pmatrix} T & \mathbf{T}^0 \\ \mathbf{0} & 0 \end{pmatrix},$$

where T is a square matrix of order n, with $T_{ii} < 0$, $T_{ij} \geqq 0$, for $i \neq j$, and such that T^{-1} exists. The n-vector $\mathbf{T}^0$ has nonnegative entries, and is equal to $-T\mathbf{e}$. The vector $\mathbf{e}$ has all entries equal to one. The vector of initial probabilities is denoted by $(\boldsymbol{\alpha}, \alpha_{n+1})$, and satisfies $\boldsymbol{\alpha}\mathbf{e} + \alpha_{n+1} = 1$, $0 \leqq \alpha_{n+1} < 1$.

The probability distribution $F(\cdot)$ of the time till absorption in the state $n+1$ is then given by

$$F(x) = 1 - \boldsymbol{\alpha} \exp(Tx)\mathbf{e} \quad \text{for } x \geqq 0.$$

The probability distribution $F(\cdot)$ is said to be of *phase-type* (in short, "F is PH"). The pair $(\boldsymbol{\alpha}, T)$ is called a *representation* of $F(\cdot)$. In this paper, we assume that $\alpha_{n+1} = 0$, so that $F(\cdot)$ does not have a jump at 0. Furthermore, we assume that the representation is such that each state has a positive probability of being visited before absorption. Under that assumption, the Markov chain with generator $T + \mathbf{T}^0 \cdot \boldsymbol{\alpha}$ is irreducible.

The moments $\mu^{(k)}$ of $F(\cdot)$ about the origin all exist and are given by

$$\mu^{(k)} = (-1)^k k! \boldsymbol{\alpha} T^{-k} \mathbf{e} \quad \text{for } k \geqq 1. \tag{1}$$

1.2 The point process. We consider N PH-distributions, with representations $(\boldsymbol{\alpha}_i, T_i)$, where T_i is a square matrix of order n_i, for $i = 1, \cdots, N$, and an N-state irreducible Markov chain with transition matrix P. If the Markov chain has made a transition to the state i, the next transition is to the state j, with probability p_{ij}, and the time between these transitions has a PH-distribution $F_i(\cdot)$, with representation $(\boldsymbol{\alpha}_i, T_i)$, independent of j. The epochs of transitions for the Markov chain correspond to the epochs of events for the point process.

We denote respectively by $N(t)$, $C(t)$ and $\Phi(t)$, the number of events in $(0, t]$, the state of the Markov chain P at time t, and the state of the Markov chain $T_{C(t)}$, at time t. In other words, suppose that the last event before t occurred at time τ. At time τ, the Markov chain P made a transition to the state $C(\tau) = j$, say, and an initial state was chosen for the Markov chain T_j, according to the probability vector $\boldsymbol{\alpha}_j$. In the interval $(\tau, t]$, the Markov chain T_j underwent zero, one, or more than one transitions, without entering its absorbing state. At time t, $C(t) = j$, and the Markov chain T_j is in the state $\Phi(t)$.

We make the following independence assumption. For every $t > 0$, the intervals of time between events are conditionally independent, given the path function of the Markov chain P. It is then clear that the process $\{N(t), C(t), \Phi(t), t \geqq 0\}$ is a Markov process with state space $\{(\nu, j, \phi); \nu \geqq 0, 1 \leqq j \leqq N, 1 \leqq \phi \leqq n_j\}$. In order to distinguish easily between the Markov chain P and the Markov chains T_i, $i = 1, \cdots, N$, we shall refer to the states of any Markov chain T_i as "*phases.*"

2. The number of events in an interval. We define the probabilities $S_{i,\xi;j,\phi}(\nu, t) = P[N(t) = \nu, C(t) = j, \Phi(t) = \phi | C(0) = i, \Phi(0) = \xi]$, and order the elements $\{(j, \phi); 1 \leqq j \leqq N, 1 \leqq \phi \leqq n_j\}$ as follows: $(1, 1), (1, 2), \cdots, (1, n_1), (2, 1), \cdots, (2, n_2), \cdots, (N, 1), \cdots, (N, n_N)$. Finally we define the block-partitioned square matrix $S(\nu, t)$

of order $n_1+n_2+\cdots+n_N$ by

$$S(\nu, t)=\begin{pmatrix} S_{1,1}(\nu, t) & S_{1,2}(\nu, t) & \cdots & S_{1,N}(\nu, t) \\ S_{2,1}(\nu, t) & S_{2,2}(\nu, t) & \cdots & S_{2,N}(\nu, t) \\ \vdots & \vdots & & \vdots \\ S_{N,1}(\nu, t) & S_{N,2}(\nu, t) & \cdots & S_{N,N}(\nu, t) \end{pmatrix},$$

where the blocks $S_{i,j}(\nu, t)$ have n_i rows and n_j columns, and the (ξ, ϕ)th element of $S_{i,j}(\nu, t)$ is equal to $S_{i,\xi;j,\phi}(\nu, t)$.

The Chapman–Kolmogorov equations for the process $\{N(t), C(t), \Phi(t), t \geqq 0\}$ may be written in matrix notation as

$$S'_{i,j}(0, t)=\begin{cases} 0, & \text{for } i \neq j, \\ S_{i,i}(0, t)T_i & \text{for } i=j, \end{cases}$$

$$S'_{i,j}(\nu, t)=S_{i,j}(\nu, t)T_j-\sum_{k=1}^{N} S_{i,k}(\nu-1, t)p_{kj}\mathbf{T}_k^0 \cdot \boldsymbol{\alpha}_j \quad \text{for } \nu \geqq 1.$$

Therefore, the matrices $S(\nu, t)$ satisfy the system of linear differential equations:

$$S'(0, t)=S(0, t)\bar{T},$$

$$S'(\nu, t)=S(\nu, t)\bar{T}-S(\nu-1, t)\bar{T}\bar{A} \quad \text{for } \nu \geqq 1,$$

with initial conditions $S(0, 0)=I$, $S(\nu, 0)=0$ for $\nu \geqq 1$, where the square matrices $\bar{T}$ and $\bar{A}$ are of order $n_1+n_2+\cdots+n_N$. The matrix $\bar{T}$ is block-diagonal and given by

$$\bar{T}=\begin{pmatrix} T_1 & 0 & 0 & \cdots & 0 \\ 0 & T_2 & 0 & \cdots & 0 \\ 0 & 0 & T_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & T_N \end{pmatrix},$$

and the block-partitioned matrix $\bar{A}$ is given by

$$\bar{A}=\begin{pmatrix} p_{11}\mathbf{e}_1 \cdot \boldsymbol{\alpha}_1 & p_{12}\mathbf{e}_1 \cdot \boldsymbol{\alpha}_2 & \cdots & p_{1N}\mathbf{e}_1 \cdot \boldsymbol{\alpha}_N \\ p_{21}\mathbf{e}_2 \cdot \boldsymbol{\alpha}_2 & p_{22}\mathbf{e}_2 \cdot \boldsymbol{\alpha}_2 & \cdots & p_{2N}\mathbf{e} \cdot \boldsymbol{\alpha}_N \\ \vdots & \vdots & & \vdots \\ p_{N1}\mathbf{e}_N \cdot \boldsymbol{\alpha}_1 & p_{N2}\mathbf{e}_N \cdot \boldsymbol{\alpha}_2 & \cdots & p_{NN}\mathbf{e}_N \cdot \boldsymbol{\alpha}_N \end{pmatrix}.$$

By $\mathbf{e}_i$ we denote an n_i-vector with each entry equal to one.

The matrix-generating function $\tilde{S}(z, t)=\sum_{\nu=0}^{\infty} z^{\nu}S(\nu, t)$, defined for $|z| \leqq 1$, satisfies the differential equation

$$\frac{\partial}{\partial t}\tilde{S}(z, t)=\tilde{S}(z, t)\bar{T}(I-z\bar{A}), \qquad \tilde{S}(z, 0)=I \quad \text{for } t \geqq 0.$$

Hence, we have that $\tilde{S}(z, t)=\exp[\bar{T}(I-z\bar{A})t]$. In particular, $\tilde{S}(0, t)=\exp(\bar{T}t)$, as is to be expected. Also, $\tilde{S}(1, t)=\exp[\bar{T}(I-\bar{A})t]$, which is again obvious, since the process $\{C(t), \Phi(t), t \geqq 0\}$ is a continuous parameter Markov chain with infinitesimal generator $\bar{T}(I-\bar{A})$.

We now define the matrix $M(t)=[(\partial/\partial z)\tilde{S}(z, t)]_{z=1}$, and the vector $\mathbf{m}(t)=M(t)\mathbf{e}$. We partition that vector as $\mathbf{m}(t)=(\mathbf{m}_1(t), \mathbf{m}_2(t), \cdots, \mathbf{m}_N(t))$, where $\mathbf{m}_i(t)$, $i=1, \cdots, N$, has n_i components. The component $m_{i,\xi}(t)$ is the expected number of events occurring

before time t, given the initial conditions $C(0)=i$, and $\Phi(0)=\xi$. Furthermore, we denote by $\boldsymbol{\gamma}$ the invariant probability vector associated with P, i.e., $\boldsymbol{\gamma}P=\boldsymbol{\gamma}$, $\boldsymbol{\gamma}\mathbf{e}=1$. Every entry of $\boldsymbol{\gamma}$ is strictly positive. Finally, we denote by $\boldsymbol{\pi}$ the stationary probability vector of $\bar{T}(I-\bar{A})$, i.e., $\boldsymbol{\pi}\bar{T}(I-\bar{A})=\mathbf{0}$, $\boldsymbol{\pi}\mathbf{e}=1$. The point process under consideration is a special case of the "versatile Markovian point process" defined in Neuts [8]. The equations (2) and (3) in the following lemma follow by adapting equation (12) of [8] to our process. The proof of the remainder of the lemma is immediate.

LEMMA 1. *The vector* $\mathbf{m}(t)$ *is given by*

$$\begin{aligned}\mathbf{m}(t) &= m^*t\mathbf{e}-(I-\Pi)[\tau^*\Pi-\bar{T}(I-\bar{A})]^{-1}\bar{T}\mathbf{e}\\ &\quad -[\Pi-\exp(\bar{T}(I-\bar{A})t)][\tau^*\Pi-\bar{T}(I-\bar{A})]^{-1}\bar{T}\mathbf{e} \quad \text{for } t\geqq 0,\end{aligned} \tag{2}$$

where the square matrix Π *of order* $n_1+n_2+\cdots+n_N$ *is equal to* $\mathbf{e}\cdot\boldsymbol{\pi}$. τ^* *is any real number such that* $\tau^*\geqq\max\{-(\bar{T}(I-\bar{A}))_{i,\phi;i,\phi};\ 1\leqq i\leqq N,\ 1\leqq\phi\leqq n_i\}$, *and* m^* *is given by*

$$m^*=-\boldsymbol{\pi}\bar{T}\mathbf{e}. \tag{3}$$

Moreover, if we partition $\boldsymbol{\pi}$ *as* $\boldsymbol{\pi}=(\boldsymbol{\pi}_1,\boldsymbol{\pi}_2,\cdots,\boldsymbol{\pi}_N)$, *where* $\boldsymbol{\pi}_i$ *is an* n_i*-vector,* $i=1,\cdots,N$, *we have that*

$$\boldsymbol{\pi}_i=c\gamma_i\boldsymbol{\alpha}_i(-T_i)^{-1} \quad \text{for } i=1,\cdots,N. \tag{4}$$

The normalizing constant c *satisfies*

$$m^*=c=(\boldsymbol{\gamma}\boldsymbol{\mu}^{(1)})^{-1}, \tag{5}$$

where the N*-vector* $\boldsymbol{\mu}^{(1)}$ *has components* $\mu_i^{(1)}=-\boldsymbol{\alpha}_iT_i^{-1}\mathbf{e}$.

Remarks. 1. The third term in (2) tends to zero as t tends to infinity, since $\Pi-\exp[\bar{T}(I-\bar{A})t]=\Pi-\exp[\tilde{S}(1,t)]$ does; therefore the first two terms give the linear asymptote of $\mathbf{m}(t)$.

2. In order to compute $\mathbf{m}(t)$, it is not necessary to evaluate the inverse of the large matrix $[\tau^*\Pi-\bar{T}(I-\bar{A})]$. It suffices to determine the vector $\mathbf{u}$ defined as

$$\mathbf{u}=-[\tau^*\Pi-\bar{T}(I-\bar{A})]^{-1}\bar{T}\mathbf{e}. \tag{6}$$

This may be done efficiently as we show in Lemma 2.

3. The main problem in computing $\mathbf{m}(t)$ from (2) therefore lies in evaluating the third term, which we denote by $\mathbf{v}(t)$. The vector $\mathbf{v}(t)$ is the solution of the system of differential equations, of order $n_1+n_2+\cdots+n_N$, given by

$$\mathbf{v}'(t)=\bar{T}(I-\bar{A})\mathbf{v}(t), \qquad \mathbf{v}(0)=(\Pi-I)\mathbf{u}.$$

LEMMA 2. *Let the vector* $\mathbf{u}$ *be partitioned as* $\mathbf{u}=(\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_N)$, *where* $\mathbf{u}_i$ *is an* n_i *vector for* $i=1,\cdots,N$. *Then*

$$\mathbf{u}_i=\tilde{\nu}\mathbf{e}+\nu_i\mathbf{e}+m^*T_i^{-1}\mathbf{e} \quad \text{for } i=1,\cdots,N, \tag{7}$$

where $\boldsymbol{\nu}=-m^*(I-P+\Gamma)^{-1}P\boldsymbol{\mu}^{(1)}$, $\tilde{\nu}=m^*/\tau^*+\frac{1}{2}m^{*2}\boldsymbol{\gamma}\boldsymbol{\mu}^{(2)}-\mathbf{h}\boldsymbol{\nu}$.

The vectors $\boldsymbol{\mu}^{(2)}$ *and* $\mathbf{h}$ *have* N *components, given by* $\mu_i^{(2)}=2\boldsymbol{\alpha}_iT_i^{-2}\mathbf{e}$, $h_i=m^*\gamma_i\mu_i^{(1)}$ *for* $i=1,\cdots N$.

Proof. The vector $\mathbf{u}$ is the unique solution to the system

$$[\tau^*\Pi-\bar{T}(I-\bar{A})]\mathbf{u}=-\bar{T}\mathbf{e}. \tag{8}$$

Upon substitution of the stated expressions for $\mathbf{u}$, it is verified that (7) indeed provides the solution to (8). The calculations, although belabored, are entirely routine; the details are omitted for the sake of brevity. □

The equation (2) now becomes

(9)

$$\mathbf{m}(t) = m^*t\mathbf{e} + m^*(I-\Pi)(\bar{T}^{-1}-R)\mathbf{e} + m^*[\Pi - \exp(\bar{T}(I-\bar{A})t)](\bar{T}^{-1}-R)\mathbf{e}, \quad \text{for } t \geqq 0,$$

where the matrix R is block-diagonal and given by

$$R = \begin{pmatrix} R_1 & 0 & \cdots & 0 \\ 0 & R_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & R_N \end{pmatrix},$$

R_i is a square matrix of order n_i, and is equal to $[(I-P+\Gamma)^{-1}P\boldsymbol{\mu}^{(1)}]_i I$.

Remark. In order to determine $(\bar{T}^{-1}-R)\mathbf{e}$, it is not necessary to invert matrices. One merely solves $N+1$ systems of linear equations, i.e., the systems $T_i\mathbf{x}_i = \mathbf{e}$, for $i = 1, \cdots, N$, and the system $(I-P+\Gamma)\mathbf{y} = P\boldsymbol{\mu}^{(1)}$.

3. A stationary version of the point process. Usually, the stationary point process, which we denote by P^*, is obtained by choosing the initial state $(C(0), \Phi(0))$ according to the probability vector $\boldsymbol{\pi}$. We consider a slightly different process, denoted by $\tilde{P}$, for which $(C(0), \Phi(0))$ is chosen by $P[C(0) = j, \Phi(0) = \xi] = \gamma_j(\boldsymbol{\alpha}_j)_\xi$. In other words, we choose the time origin so that at time 0-, an event has occurred, the type of the next interval is chosen according to the stationary vector $\boldsymbol{\gamma}$ of P. In view of our ultimate objective of using this process to model arrivals to queueing systems, the process $\tilde{P}$ has the following interesting property.

Let us denote by X_n the interval of time between the $(n-1)$st and the nth event (between time 0 and the first event, if $n = 1$). The following result is elementary.

LEMMA 3. *The random variables* $\{X_n, n \geqq 1\}$ *have a common marginal distribution* $r(\cdot)$. *The distribution* $r(\cdot)$ *is* PH, *and has a representation* $(\mathbf{a}, \bar{T})$, *where the vector* $\mathbf{a}$ *has* $n_1 + n_2 + \cdots n_N$ *components and is given by* $\mathbf{a} = (\gamma_1\boldsymbol{\alpha}_1, \gamma_2\boldsymbol{\alpha}_2, \cdots, \gamma_N\boldsymbol{\alpha}_N)$. *Therefore*

$$r(x) = 1 - \sum_{i=1}^{N} \gamma_i\boldsymbol{\alpha}_i \exp(T_i x)\mathbf{e} \quad \textit{for } x \geqq 0,$$

and the k-th moment $m^{(k)}$ *of* $r(\cdot)$ *about the origin is equal to*

$$m^{(k)} = \boldsymbol{\gamma}\boldsymbol{\mu}^{(k)} \quad \textit{for } k \geqq 1.$$

The N-vectors $\boldsymbol{\mu}^{(k)}$ *have entries* $\mu_i^{(k)} = (-1)^k k!\boldsymbol{\alpha}_i T_i^{-k}\mathbf{e}$.

In the remainder of this section, we examine the correlation structure of the intervals of time between events. This is related to recent work by Simon [12]. Because of the difference in our approach, we postpone discussion of this relation to the end of this section.

We now introduce the notion of *linear dependence* for PH-distributions.

DEFINITION 1. The set $\{(\boldsymbol{\alpha}_i, T_i), 1 \leqq i \leqq N\}$ is a set of *linearly independent* PH-distributions if and only if $\sum_{i=1}^{N} d_i F_i(x) = 0$ for all $x \geqq 0$ implies that d_i is equal to zero for $i = 1, \cdots, N$, where $F_i(x) = 1 - \boldsymbol{\alpha}_i \exp(T_i x)\mathbf{e}$ for $x \geqq 0$.

DEFINITION 2. The PH-distribution $(\boldsymbol{\beta}, B)$ is a *linear combination* of the PH-distributions $(\boldsymbol{\alpha}_i, T_i)$, $i = 1, \cdots, N$, if and only if there exist $\{d_1, d_2, \cdots, d_N\}$, such that $d_i \neq 0$ for some i, and $1 - \boldsymbol{\beta}\exp(Bx)\mathbf{e} = \sum_{i=1}^{N} d_i F_i(x)$ for all $x \geqq 0$, where the $F_i(x)$ are defined above.

It is clear that for any such set $\{d_1, \cdots, d_N\}$ we have $\sum_{i=1}^{N} d_i = 1$.

Remarks. 1. The term "linearly independent" has been chosen for the following reason. We easily observe that $\sum_{i=1}^{N} d_i F_i(x)$ is equal to zero for all positive x if and only if $\sum_{i=1}^{N} d_i \boldsymbol{\alpha}_i T_i^k \mathbf{e}$ is equal to zero for all $k \geqq 1$. Definition 1 is therefore equivalent to the condition that the infinite vectors $(\boldsymbol{\alpha}_i T_i \mathbf{e}, \boldsymbol{\alpha}_i T_i^2 \mathbf{e}, \cdots)$, $i = 1, \cdots, N$, are linearly independent.

2. If $(\boldsymbol{\beta}, B)$ is a linear combination of $\{(\boldsymbol{\alpha}_i, T_i), i = 1, \cdots, N\}$, and $d_i \geqq 0$, for all $i = 1, \cdots, N$, then $(\boldsymbol{\beta}, B)$ and $(\boldsymbol{\beta}_0, B_0)$ are two representations of the same PH-distribution, where

$$\boldsymbol{\beta}_0 = (d_1 \boldsymbol{\alpha}_1, d_2 \boldsymbol{\alpha}_2, \cdots, d_N \boldsymbol{\alpha}_N),$$

and

$$B_0 = \begin{pmatrix} T_1 & 0 & \cdots & 0 \\ 0 & T_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & T_N \end{pmatrix}.$$

The proof is elementary. Similarly, if $d_i < 0$ for $i = 1, \cdots, J$ and $d_i \geqq 0$ for $i = J+1, \cdots, N$, then clearly

$$1 + \sum_{i=1}^{J} |d_i| = \sum_{i=J+1}^{N} d_i.$$

If we set the latter quantity equal to $\tilde{d}$, then $(\boldsymbol{\beta}_1, B_1)$ and $(\boldsymbol{\beta}_2, B_2)$, where

$$\boldsymbol{\beta}_1 = \left(\frac{1}{\tilde{d}} \boldsymbol{\beta}, \frac{|d_1|}{\tilde{d}} \boldsymbol{\alpha}_1, \cdots, \frac{|d_J|}{\tilde{d}} \boldsymbol{\alpha}_J\right), \qquad B_1 = \begin{pmatrix} B & 0 & \cdots & 0 \\ 0 & T_1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & T_J \end{pmatrix},$$

$$\boldsymbol{\beta}_2 = \left(\frac{d_{J+1}}{\tilde{d}} \boldsymbol{\alpha}_{J+1}, \cdots, \frac{d_N}{\tilde{d}} \boldsymbol{\alpha}_N\right), \qquad B_2 = \begin{pmatrix} T_{J+1} & 0 & \cdots & 0 \\ 0 & T_{J+2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & T_N \end{pmatrix},$$

are two representations of the same PH-distribution.

Now let the PH-distributions $\{(\boldsymbol{\alpha}_i, T_i), i = 1, \cdots, N\}$ which define the point process be expressed as linear combinations of *linearly independent* PH-distributions $\{(\boldsymbol{\zeta}_j, Z_j), j = 1 \cdots L\}$, with $L \leqq N$, i.e.,

$$F_i(x) = \sum_{j=1}^{L} d_{ij} G_j(x) \quad \text{for all } x \geqq 0 \quad i = 1, \cdots, N,$$

where $G_j(x) = 1 - \boldsymbol{\zeta}_j \exp(Z_j x)\mathbf{e}$. We denote by D the matrix with (i, j)th element equal to d_{ij}. We also define the vectors $\mathbf{F}(x) = (F_1(x), F_2(x), \cdots, F_N(x))$, $\mathbf{G}(x) = (G_1(x), G_2(x), \cdots, G_L(x))$, and $\boldsymbol{\mu}^{*(k)} = (\mu_1^{*(k)}, \cdots, \mu_L^{*(k)})$, for $k \geqq 1$, where $\mu_j^{*(k)}$ is the kth moment about the origin of $G_j(\cdot)$.

We then clearly have that

$$\mathbf{F}(x) = D\mathbf{G}(x) \quad \text{for } x \geqq 0, \tag{10}$$

and

$$\boldsymbol{\mu}^{(k)} = D\boldsymbol{\mu}^{*(k)} \quad \text{for } k \geqq 1. \tag{11}$$

It is also clear that $D\mathbf{e} = \mathbf{e}$.

We emphasize that the set of PH-distributions $\{(\zeta_j, Z_j), j = 1, \cdots, L\}$ is not uniquely determined, and is not necessarily a subset of $\{(\alpha_i, T_i), i = 1, \cdots, N\}$. The following is an illustrative example.

If

$$(\alpha_1, T_1) = \left\{(\tfrac{1}{2}, \tfrac{1}{2}), \begin{pmatrix} -\lambda_1 & 0 \\ 0 & -\lambda_2 \end{pmatrix}\right\},$$

$$(\alpha_2, T_2) = \left\{(\tfrac{2}{3}, \tfrac{1}{3}), \begin{pmatrix} -\lambda_1 & 0 \\ 0 & -\lambda_3 \end{pmatrix}\right\},$$

$$(\alpha_3, T_3) = \left\{(\tfrac{1}{3}, \tfrac{2}{3}), \begin{pmatrix} -\lambda_2 & 0 \\ 0 & -\lambda_3 \end{pmatrix}\right\},$$

$$(\alpha_4, T_4) = \left\{(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}), \begin{pmatrix} -\lambda_1 & 0 & 0 \\ 0 & -\lambda_2 & 0 \\ 0 & 0 & -\lambda_3 \end{pmatrix}\right\},$$

then we may either choose

$$(\zeta_j, Z_j) = (\alpha_j, T_j), \qquad j = 1, 2, 3,$$

with

$$D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{2}{5} & \frac{1}{5} & \frac{2}{5} \end{pmatrix},$$

or alternatively

$$(\zeta_j, Z_j) = \{(1), (-\lambda_j)\}, \qquad j = 1, 2, 3,$$

in which case D is given by

$$D = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}.$$

The next result follows easily from Lemma 3.

LEMMA 4. *The covariance between X_n and X_{n+m} for $n \geqq 0$, $m \geqq 1$ is given by*

$$\text{Cov}\,(X_n, X_{n+m}) = \boldsymbol{\mu}^{(1)} \Delta(\boldsymbol{\gamma})(P^m - \Gamma)\boldsymbol{\mu}^{(1)} \tag{12}$$

$$= \boldsymbol{\mu}^{*(1)} D^T \Delta(\boldsymbol{\gamma})(P^m - \Gamma) D \boldsymbol{\mu}^{*(1)}, \tag{13}$$

where for any vector $\mathbf{x}$, *the matrix* $\Delta(\mathbf{x})$ *is defined by* diag $(x_1, \cdots, x_N)$.

From Lemmas 3 and 4, we conclude that, for the process $\tilde{P}$, the intervals of time between events are identically distributed and are in general correlated. If the Markov chain P is aperiodic, then $\lim_{m\to\infty} P^m = \Gamma$, and the covariance of X_n and X_{n+m} tends to zero as m tends to infinity.

To conclude this section, we shall now examine under what condition the process P is a renewal process. The next theorem holds for any set $\{G_1(\cdot), \cdots, G_L(\cdot)\}$ of linearly independent PH-distributions which satisfy (10).

THEOREM 1. *The process $\tilde{P}$ is a renewal process if and only if the following property holds.*

For all $k \geqq 2$, and for all $(\tau_1, \cdots, \tau_k)$ satisfying $1 \leqq \tau_i \leqq L$ for $i = 1, \cdots, k$,

$$\boldsymbol{\gamma}\Delta_{\tau_1} \prod_{i=2}^{k} [(P-\Gamma)\Delta_{\tau_i}]\mathbf{e} = 0, \tag{14}$$

or, equivalently,

$$\boldsymbol{\gamma}\Delta_{\tau_1}\Big(\prod_{i=2}^{k} P\Delta_{\tau_i}\Big)\mathbf{e} = \prod_{i=1}^{k} (\boldsymbol{\gamma}\mathbf{D}_{\cdot\tau_i}), \tag{15}$$

where $\mathbf{D}_{\cdot j}$ represents the jth column of the matrix D, and $\Delta_j = \operatorname{diag}(\mathbf{D}_{\cdot j})$.

Proof. Since the variables $\{X_n, n \geqq 0\}$ are identically distributed, $\tilde{P}$ is a renewal process if and only if for all $k \geqq 2$, the random variables $X_1, X_2, \cdots, X_k$ are independent, which is true if and only if

$$P\Big[\bigcap_{i=1}^{k} \{X_i \leqq x_i\}\Big] = \prod_{i=0}^{k} P[X_i \leqq x_i] \quad \text{for all } x_1, \cdots, x_k \geqq 0;$$

equivalently, if

$$\sum_{\substack{\nu_i=1 \\ 1\leqq i\leqq k}}^{N} \{\gamma_{\nu_i} F_{\nu_i}(x_1) \prod_{j=2}^{k} P_{\nu_{j-1}\nu_j} F_{\nu_j}(x_j) - \prod_{j=1}^{k} \gamma_{\nu_j} F_{\nu_j}(x_j)\} = 0,$$

for all $x_1, \cdots, x_k \geqq 0$, if and only if (by (10))

$$\sum_{\substack{\nu_i=1 \\ 1\leqq i\leqq k}}^{N} \sum_{\tau_i=1}^{L} \Big\{\gamma_{\nu_i} d_{\nu_i\tau_i}\Big[\prod_{j=2}^{k} (P_{\nu_{j-1}\nu_j} - \gamma_{\nu_j}) d_{\nu_j\tau_j}\Big] \prod_{j=1}^{k} G_{\tau_j}(x_j)\Big\} = 0 \quad \text{for all } x_1, \cdots, x_k \geqq 0.$$

Since the PH-distributions $G_i(\cdot)$ are linearly independent, this holds if and only if

$$\sum_{\substack{\nu_i=1 \\ 1\leqq i\leqq k}}^{N} \Big\{\gamma_{\nu_i} d_{\nu_i\tau_i}\Big[\prod_{j=2}^{k} (P_{\nu_{j-1}\nu_j} - \gamma_{\nu_j}) d_{\nu_j\tau_j}\Big]\Big\} = 0,$$

for all $\tau_1, \cdots, \tau_k$ such that $1 \leqq \tau_i \leqq L$. The condition (14) is now obvious, and it is a simple matter to prove that (14) and (15) are equivalent. □

This theorem provides us with a technical condition which is not very attractive. The following corollary is more interesting and useful for modeling purposes.

COROLLARY 1. *For $\tilde{P}$ to be a renewal process, it is sufficient that*

$$(P-\Gamma)D = 0, \tag{16}$$

or that

$$D^T \Delta(\boldsymbol{\gamma})(P-\Gamma) = 0. \tag{17}$$

If $L = 1$, then both conditions are always satisfied and $\tilde{P}$ is always a renewal process. If $L = N$, then both conditions are necessary, as $\tilde{P}$ is a renewal process if and only if the matrix $P-\Gamma$ is equal to zero.

Proof. The condition (16) is obviously sufficient, as $(P-\Gamma)\mathbf{D}_{\cdot i} = \mathbf{0}$ for all i implies that $(P^n-\Gamma)\Delta_i\mathbf{e} = P^{n-1}(P-\Gamma)\mathbf{D}_{\cdot i} = \mathbf{0}$ for each $n \geqq 1$ and each $i = 1, \cdots, L$, which in turn implies (14).

Similarly, the condition (17) is sufficient, since

$$\boldsymbol{\gamma}\Delta_i(P^n-\Gamma)=\boldsymbol{\gamma}\Delta_i(P-\Gamma)P^{n-1}=\mathbf{D}^T{}_i\Delta(\boldsymbol{\gamma})(P-\Gamma)P^{n-1}.$$

If L is equal to one, then $D=\mathbf{e}$ and both conditions (16) and (17) are satisfied.

If L is equal to N, then D may be chosen equal to I, and both (16) and (17) reduce to $P-\Gamma=0$. The necessary part of the condition results from (14): If $k=1$, then $\boldsymbol{\gamma}\Delta_i(P-\Gamma)\Delta_j\mathbf{e}$ must be equal to zero for all $i, j=1,\cdots,N$. As $D=I$, this reduces to $\gamma_i(P_{ij}-\gamma_j)=0$ for all $i, j=1,\cdots,N$. Since $\gamma_i>0$ for all i, Corollary 1 is proved. □

Remark. One easily proves that condition (16) holds if and only if there exist a vector $\mathbf{v}$ such that $PD=\mathbf{e}\cdot\mathbf{v}$. Therefore, it is not necessary to determine $\boldsymbol{\gamma}$ in order to check whether (16) holds or not. This condition is most easily interpreted when the entries of D are equal to zero or one. In that case, (16) implies two consequences:

(a). The Markov chain P is lumpable to a Markov chain P' on $\{1,\cdots,L\}$.

(b) The rows of P' are all identical.

In other words, the PH semi-Markov process is in fact a renewal process, the distribution of each interval being a mixture of the linearly independent distribution $\{G_i(\cdot), i=1,\cdots,L\}$.

As we have observed, the entries of D may take any real value. If those values were all positive, one might still interpret (16) as implying that the Markov chain is lumpable in some randomized way (since the row sums of D are equal to one), to a Markov chain P' such that all the rows of P' are identical. This interpretation appears difficult to extend to the case where D contains negative entries.

The condition (17) is more difficult to interpret. After examining the case where the entries of D are zero or one only, we have tentatively reached the following conclusion. If the relation (17) holds, then the Markov chain P and the mixture of distributions $\{G_i(\cdot), i=1,\cdots,L\}$ generated by D are such that, starting with the initial probability vector $\boldsymbol{\gamma}$, the semi-Markov process is completely randomized and becomes a renewal process.

We show in the following example that the conditions (16) and (17) are not equivalent. The matrix P and the vector $\boldsymbol{\gamma}$ are given by

$$P=\begin{bmatrix}\frac{1}{2}&\frac{1}{2}&0\\ \frac{1}{4}&\frac{1}{4}&\frac{1}{2}\\ \frac{1}{4}&\frac{1}{4}&\frac{1}{2}\end{bmatrix},\qquad \boldsymbol{\gamma}=(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}).$$

If

$$(\boldsymbol{\alpha}_1, T_1)=\left\{(\tfrac{1}{2},\tfrac{1}{2}),\begin{bmatrix}-\lambda_1&0\\0&-\lambda_2\end{bmatrix}\right\},$$

$$(\boldsymbol{\alpha}_2, T_2)=\{(1),(-\lambda_2)\},$$

$$(\boldsymbol{\alpha}_3, T_3)=\left\{(\tfrac{1}{4},\tfrac{1}{3}),\begin{bmatrix}-\lambda_1&0\\0&-\lambda_2\end{bmatrix}\right\},$$

then we may choose $G_1(\cdot)$ and $G_2(\cdot)$ to be exponential distributions with parameters λ_1 and λ_2 respectively, and we have that

$$D=\begin{bmatrix}\frac{1}{2}&\frac{1}{2}\\ 0&1\\ \frac{1}{4}&\frac{3}{4}\end{bmatrix},$$

and one easily verifies that (16) holds, but not (17). If we now change $(\boldsymbol{\alpha}_3, T_3)$ to

$\{(1), (-\lambda_1)\}$, then the matrix D becomes

$$D = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

and (16) does not hold, while (17) is satisfied.

Remark. If neither (16) nor (17) holds, the combined condition,

$$D^T \Delta(\gamma)(P - \Gamma)D = 0 \tag{18}$$

is not a sufficient condition for $\tilde{P}$ to be a renewal process (as we show below in one example), but is the necessary and sufficient condition for two *successive* intervals of time to be independent. Consider the four distributions

$$(\alpha_1, T_1) = \{(1), [-\lambda_1]\}, \qquad (\alpha_2, T_2) = \left\{(\tfrac{1}{3}, \tfrac{2}{3}), \begin{bmatrix} -\lambda_1 & 0 \\ 0 & -\lambda_2 \end{bmatrix}\right\},$$

$$(\alpha_3, T_3) = \{(1), [-\lambda_2]\}, \qquad (\alpha_4, T_4) = \left\{(\tfrac{2}{3}, \tfrac{1}{3}), \begin{bmatrix} -\lambda_1 & 0 \\ 0 & -\lambda_2 \end{bmatrix}\right\}.$$

The matrix D is given by

$$D = \begin{bmatrix} 1 & 0 \\ \frac{1}{3} & \frac{2}{3} \\ 0 & 1 \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix}.$$

If the matrix P is chosen to be

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix},$$

then $\gamma = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and one can check that the relation (14) holds for $k = 2$, but not for $k \geqq 3$. In fact, we observe upon closer examination of the process, that X_n and X_{n+m} are independent random variables if and only if $m = 2k + 1$ for some integer k.

Remark. It is possible to strengthen this corollary under special additional conditions. We shall present these results in the next section, together with the examples in which these special conditions arise.

Simon [12] examines equivalences for Markov-renewal processes and, in particular, the conditions under which a Markov-renewal process is equivalent to a renewal process. It appears that for our process $\tilde{P}$, Theorems 2.2.1, 2.2.2 and 2.2.9 of [12] respectively correspond to the sufficient condition (17), and to the cases $L = 1$ and $L = N$ in Corollary 1. Because of the special structure of our point process, and in particular because we consider PH-distributions, we have obtained conditions on constant matrices, while Simon obtains conditions on matrices of functions, which have to be examined for all values for the argument of these functions. We shall not attempt to make a more detailed comparison in this short space.

4. Examples.

4.1 Exponential distributions. The case where the PH-distributions (α_i, T_i) are exponential, respectively with parameters λ_i, is particularly simple. We then immedi-

ately obtain that

$$\bar{T} = -\Delta(\boldsymbol{\lambda}), \quad \bar{A} = P, \quad m^* = (\boldsymbol{\gamma}\boldsymbol{\lambda}^{-1})^{-1},$$
$$\pi_i = m^* \gamma_i \lambda_i^{-1} \quad \text{for } i = 1, \cdots, N,$$

where the vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^{-1}$ respectively have components equal to λ_i and λ_i^{-1}. After some simple manipulations, (9) reduces to

$$m(t) = m^* t\mathbf{e} + m^*(I - m^*\Gamma\Delta(\boldsymbol{\lambda}^{-1}))(I - P + \Gamma)^{-1}\boldsymbol{\lambda}^{-1} + m^*[m^*\Gamma\Delta(\boldsymbol{\lambda}^{-1}) - \exp(-\Delta(\boldsymbol{\lambda})(I-P)t)](I - P + \Gamma)^{-1}\boldsymbol{\lambda}^{-1}. \tag{19}$$

Exponential distributions with different parameters are linearly independent. Therefore, the distributions $\{(\zeta_j, Z_j), 1 \leqq j \leqq L\}$ may be chosen to be the set of different exponential distributions in $\{(\boldsymbol{\alpha}_i, T_i), i = 1, \cdots, N\}$, and the matrix D has a very simple structure: Each element of D is equal to either zero or one, each row of D contains exactly one element equal to one, each column of D contains at least one element equal to one. We may then strengthen Corollary 1 as follows. The technical proof is belabored. We do not reproduce it here since it cannot be extended to $L \leqq N-2$, and, therefore, is of little interest.

COROLLARY 1. *If the entries of D are each equal to zero or one, and if $L = N-1$, then a necessary and sufficient condition for $\tilde{P}$ to be a renewal process is that at least one of* (16) *or* (17) *holds.*

If the matrix P has identical rows, then $\tilde{P}$ is obviously a renewal process, with hyperexponential intervals between events, and the matrix $(I - P + \Gamma)^{-1}$ in (9) may be replaced by the identity matrix.

4.2 Platooned events. Let us assume that the process consists of groups of events, the number of events in a group has a discrete PH-distribution $(\mathbf{f}, F)$, the intervals of time between events in a given group have PH-distribution $(\boldsymbol{\alpha}, A)$, while the intervals of times between groups have a PH-distribution $(\boldsymbol{\beta}, B)$. Such a process may be used to model platooned arrivals to a system, as is done in Neuts and Chakravarthy [9]. Then,

$$P = \left(\begin{array}{c|c} F & \mathbf{e} - F\mathbf{e} \\ \hline \mathbf{f}F & 1 - \mathbf{f}F\mathbf{e} \end{array} \right), \tag{20}$$

P is a square matrix of order N if F is a matrix of order $N-1$; $(\boldsymbol{\alpha}_i, T_i) = (\boldsymbol{\alpha}, A)$ for $i = 1, \cdots, N-1$, and $(\boldsymbol{\alpha}_N, T_N) = (\boldsymbol{\beta}, B)$.

Simple calculations yield that the stationary probability vector $\boldsymbol{\gamma}$ of P is given by $\boldsymbol{\gamma} = (\tilde{\boldsymbol{\gamma}}, \gamma_N)$, where

$$\tilde{\boldsymbol{\gamma}} = \mathbf{v}F, \qquad \gamma_N = 1 - \tilde{\boldsymbol{\gamma}}\mathbf{e}, \tag{21}$$

$$\mathbf{v} = (\mathbf{f}(I-F)^{-1}\mathbf{e})^{-1}\mathbf{f}(I-F)^{-1}, \tag{22}$$

and that

$$m^* = [(\tilde{\boldsymbol{\gamma}}\mathbf{e})(-\boldsymbol{\alpha}A^{-1}\mathbf{e}) + \gamma_N(-\boldsymbol{\beta}B^{-1}\mathbf{e})]^{-1},$$
$$\boldsymbol{\pi}_i = \begin{cases} -m^*\tilde{\gamma}_i\boldsymbol{\alpha}A^{-1} & \text{for } i = 1, \cdots, N-1, \\ -m^*\gamma_N\boldsymbol{\beta}B^{-1} & \text{for } i = N. \end{cases}$$

The expression (9) does not simplify much.

If the PH-distributions $(\boldsymbol{\alpha}, A)$ and $(\boldsymbol{\beta}, B)$ are different, then we clearly may choose D equal to

$$D = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{23}$$

and we may complete Corollary 1 as follows.

COROLLARY 1″. *If the matrix D is given by* (23), *then the process $\tilde{P}$ is a renewal process if and only if*

$$^{T}D\Delta(\boldsymbol{\gamma})(P^{n}-\Gamma)D=0 \quad \textit{for all } n \geqq 1. \tag{24}$$

Proof. We partition the matrix P and the vector $\boldsymbol{\gamma}$ as

$$P=\left(\begin{array}{c|c} P_{11} & \mathbf{P}_{12} \\ \hline \mathbf{P}_{21} & P_{22} \end{array}\right), \qquad \boldsymbol{\gamma}=(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2).$$

Simple calculations yield that

$$\boldsymbol{\gamma}_1=\boldsymbol{\gamma}_2\mathbf{P}_{21}(I-P_{11})^{-1}, \qquad \boldsymbol{\gamma}_2=(1+\mathbf{P}_{21}(I-P_{11})^{-1}\mathbf{e})^{-1}.$$

(a). Equation (24) is a necessary condition. For $n=1$, it is equivalent to (18), the condition for two successive intervals to be independent. For $n=2$, we consider (14) with $k=3$. Since $\boldsymbol{\gamma}\Delta_{\tau_1}(P-\Gamma)\Delta_1(P-\Gamma)\Delta_{\tau_3}\mathbf{e}+\boldsymbol{\gamma}\Delta_{\tau_1}(P-\Gamma)\Delta_2(P-\Gamma)\Delta_{\tau_3}\mathbf{e}=\boldsymbol{\gamma}\Delta_{\tau_1}(P-\Gamma)^2\Delta_{\tau_3}\mathbf{e}$ must be equal to zero, this proves (24) for $n=2$. Similarly, we prove that (24) is necessary for larger values of n.

(b) Equation (24) is a sufficient condition. We prove by recurrence that (24) implies that

$$\boldsymbol{\gamma}_1 C_1^{n+1}\mathbf{e}=(\mathbf{P}_{21}-\boldsymbol{\gamma}_1)C_1^{n}\mathbf{e}=0 \quad \text{for all } n \geqq 0 \tag{25}$$

and

$$(P-\Gamma)^n=\left[\begin{array}{c|c} C_1^n-\sum_{\nu=1}^{n-1} C_1^{\nu}\mathbf{e}\cdot(\mathbf{P}_{21}-\boldsymbol{\gamma}_1)C_1^{m-1-\nu} & -C_1^n\mathbf{e} \\ \hline (\mathbf{P}_{21}-\boldsymbol{\gamma}_1)C_1^{n-1} & 0 \end{array}\right],$$

where $C_1=P_{11}-\mathbf{e}\cdot\boldsymbol{\gamma}_1$.

It is then a simple matter to verify that for ν equal either to 1 or 2, $\Delta_2(P-\Gamma)^m\Delta_\nu\mathbf{e}=\mathbf{0}$, and $(P-\Gamma)^n\Delta_1(P-\Gamma)^m\Delta_\nu\mathbf{e}=(P-\Gamma)^{n+m}\Delta_\nu\mathbf{e}$ for $n, m \geqq 1$. Therefore, the left-hand side of (14) is equal to zero if $\tau_i=2$ for any $i=2, \cdots, k-1$ and is otherwise equal to

$$(^{T}D\Delta(\boldsymbol{\gamma})(P-\Gamma)^{k-1}D)_{\tau_1,\tau_k}.$$

This completes the proof. □

Remark. This corollary has the following simple interpretation. If D is given by (23), then $N-1$ of the distributions $\{(\boldsymbol{\alpha}_i, T_i), i=1, \cdots, N\}$ have a "common" type; the last one has an "odd" type. Let N_1 denote the number of intervals of the common type between two consecutive intervals of the odd type, and let N_2 similarly denote the number of intervals of the odd type between two intervals of the common type. The following proposition results from Corollary 1″.

PROPOSITION. *The process P is a renewal process if and only if N_1 and N_2 both have a geometric distribution, the parameters being respectively equal to* $(1-P_{22})$ *and* P_{22}.

Proof. Clearly, N_2 has a geometric distribution with parameter P_{22}. The equations (25) may be written as

$$\mathbf{P}_{21}P_{11}^{n}\mathbf{e} = (\boldsymbol{\gamma}_1\mathbf{e})^{n+1} \quad \text{for } n \geqq 0. \tag{26}$$

As $P[N_1 = n] = \mathbf{P}_{21}(P_{11}^{n-1} - P_{11}^{n})\mathbf{e} = (\boldsymbol{\gamma}_1\mathbf{e})^{n}(1 - \boldsymbol{\gamma}_1\mathbf{e}) = (1-\gamma_2)^{n}\gamma_2 = (1-P_{22})^{n}P_{22}$, this completes the proof. □

In the present case, from (20)–(22) and (26), it results that the process of platooned events is a renewal process if and only if

$$\mathbf{f}F^{n}\mathbf{e} = (1 - (\mathbf{f}(I-F)^{-1}\mathbf{e})^{-1})^{n} \quad \text{for } n \geqq 0,$$

in other words, if the number of events in a platoon, after the first one, has a geometric distribution with parameter $1 - [\mathbf{f}(I-F)^{-1}\mathbf{e}]^{-1}$.

4.3 The interrupted Poisson process. This process is used in models for telephone engineering (Heffes [3]). We consider a process in a random environment, with two alternating environment states. Both states have exponential duration, with parameters σ_1 and σ_2 respectively. While the process is in the first environment state (on-state) a Poisson process of rate λ is turned on; in the second state (off-state), no arrivals can occur. In fact, it appears that the interrupted Poisson process is a very special case of the type of processes analysed in the present paper. Neuts and Chakravarthy [9] have shown that the interrupted Poisson process is a special case of the platooned events process, and that the number of events in a platoon is geometric; therefore, the stationary interrupted Poisson process is a renewal process. It is observed in [9] that the process can be described by two states and two PH-distributions, the matrix P and the PH-distributions being as follows.

$$P = \begin{bmatrix} \dfrac{\lambda}{\lambda+\sigma_1} & \dfrac{\sigma_1}{\lambda+\sigma_1} \\ \dfrac{\lambda}{\lambda+\sigma_1} & \dfrac{\sigma}{\lambda+\sigma_1} \end{bmatrix}, \tag{27}$$

$$F_1(\cdot) \equiv (\boldsymbol{\alpha}_1, T_1) = \{(1), [-(\lambda+\sigma_1)]\}, \tag{28}$$

$$F_2(\cdot) \equiv (\boldsymbol{\alpha}_2, T_2) = \left\{(1,0,0), \begin{bmatrix} -(\lambda+\sigma_1) & \lambda+\sigma_1 & 0 \\ \sigma_2\sigma_1(\lambda+\sigma_1)^{-1} & -\sigma_2 & \sigma_2\lambda(\lambda+\sigma_1)^{-1} \\ 0 & 0 & -(\lambda+\sigma_1) \end{bmatrix}\right\} \tag{29}$$

The first state of the Markov chain P corresponds to the following event {the interrupted Poisson process is in the on-state, and an arrival will occur before the end of the on-state}. The second state corresponds to the following composite event {the interrupted Poisson process is in the on-state and no arrivals will occur before the next off-state *or* the process is in the off-state, *or* the process has returned to the on-state and an arrival will occur before the next off-state}. From the structure (27) of P, it is obvious that the interrupted Poisson process is a renewal process. Kuczura [5] has already shown, by different methods, that the interrupted Poisson process is a renewal process with a hyperexponential interval distribution. We have shown that the interval distribution $G(\cdot)$ is the following mixture of the distributions $F_1(\cdot)$ and $F_2(\cdot)$, (Equations (28) and (29)): $G(\cdot) = \lambda(\lambda+\sigma_1)^{-1}F_1(\cdot) + \sigma_1(\lambda+\sigma_1)^{-1}F_2(\cdot)$. To verify that $G(\cdot)$ is a representation for a hyperexponential distribution, one merely determines its Laplace–Stieltjes transform.

REFERENCES

[1] E. CINLAR, *Markov renewal theory*, Adv. in Appl. Prob., 1 (1969), pp. 123–187.
[2] ———, *Introduction to Stochastic Processes*, Prentice-Hall, Englewood Cliffs, NJ, 1975.
[3] H. HEFFES, *Analysis of first-come, first-served queueing systems with peaked inputs*, Bell Syst. Tech. J. 52 (1973), pp. 1215–1228.
[4] J. KEMENY AND L. SNELL, *Finite Markov Chains*, Van Nostrand, Princeton, NJ, 1960.
[5] A. KUCZURA, *The interrupted Poisson process as an overflow process*, Bell Syst. Tech. J., 52 (1973), pp. 437–448.
[6] G. LATOUCHE, *On a Markovian queue with weakly correlated interarrival times*, J. Appl. Prob., 18 (1981), pp. 190–203.
[7] M. F. NEUTS, *Probability distributions of phase type*, in Liber Amicorum Prof. Emeritus H. Florin, Dept. of Math., Univ. Louvain, Belgium, 1975, pp. 173–206.
[8] ———, *A versatile Markovian point process*, J. Appl. Prob., 16 (1979), pp. 764–779.
[9] M. F. NEUTS AND S. CHAKRAVARTHY, *A single server queue with platooned arrivals and phase type services*, Tech. Rep. 50B, Applied Mathematics Institute, Univ. of Delaware, 1980. European J. Oper. Res., to appear.
[10] R. PYKE, *Markov renewal processes: Definition and preliminary properties*, Ann. Math. Stat., 32 (1961), pp. 1231–1242.
[11] ———, *Markov renewal processes with finitely many states*, Ann. Math. Stat., 32 (1961), pp. 1243–1259.
[12] B. SIMON, *Equivalent Markov-renewal processes*, Tech. Rep. VTR 8001, Dept. of Industrial Engineering and Operations Research, Virginia Polytechnic Institute and State University, Blacksburg, 1979.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 1, March 1982

0196-5212/82/0301-0009 $01.00/0

ON GENERIC RIGIDITY IN THE PLANE*

L. LOVÁSZ† AND Y. YEMINI‡

Abstract. Let G be a graph. Let us place the points of G in "general" position in the plane and then replace its edges by rigid bars (with flexible joints). We would like to know if the resulting structure is rigid and if not, compute its "degree of freedom". This problem was solved by Laman [6] (see also [2]). In this note we give some new formulations and a new proof of Laman's theorem, based on matroid theory, and then apply these to prove the following result: if G is 6-connected, then it will be rigid in the plane. We also construct infinitely many 5-connected graphs which do not have this property.

1. Definitions. Let G be a graph whose vertices are points in the plane. Such a graph will be called a *plane structure*. (We shall visualize its edges as rigid bars.) A plane structure is *generic* if the coordinates of its points are algebraically independent over the rational field. (This highly nonmechanical assumption means that there is no "degeneracy" in the position of the points; it will be used in the form that certain polynomials of the coordinates do not vanish, and therefore all conclusions below are also valid for structures obtained by displacing the points of G a little but arbitrarily.)

An *infinitesimal motion* of G is an assignment of a plane vector $v(x)$ to every vertex x such that

$$(v(x)-v(y))\cdot(x-y)=0 \tag{1}$$

for every edge (x, y) of G (if $v(x)$ is viewed as a velocity of the point x then this condition means that no edge is compressed or stretched, at least momentarily). A *mechanical motion* is a parametrized family $(G_t: a \leqq t \leqq b)$ of plane structures, all embeddings of the same graph G, such that the position $x(t)$ of each point of G is a differentiable function of t and

$$|x(t)-y(t)| = \text{constant} \tag{2}$$

for every edge (x, y) of G. By squaring and differentiating (2) we get that

$$(x(t)-y(t))(\dot{x}(t)-\dot{y}(t))=0,$$

i.e., for every t, the vectors $v(x)=\dot{x}(t)$ define an infinitesimal motion of G_t. In general, if we consider a structure G and an infinitesimal motion of G, then this does not necessarily arise from a mechanical motion; but if G is generic then every infinitesimal motion of G is the velocity of some mechanical motion.

The infinitesimal motions of G form a linear space (with respect to pointwise addition and multiplication by scalars). The rigid motions of G yield a 3-dimensional subspace of this linear space. The codimension of this subspace of rigid motions in the space of all infinitesimal motions is called the *degree of freedom* of G. The degree of freedom of G will be denoted by $f(G)$. The structure G is called *rigid* if $f(G)=0$. Observe that if G has n points then

$$0 \leqq f(G) \leqq 2n-3.$$

* Received by the editors April 22, 1981.

† József Attila University, Szeged, Hungary.
‡ Information Sciences Institute, University of Southern California, Marina del Rey, California 90291.

Let $V(G)=\{v_1, \cdots, v_n\}$, $E(G)=\{e_1, \cdots, e_m\}$, and let us orient every edge arbitrarily. Set

$$a_{ij}=\begin{cases} 1 & \text{if } x_i \text{ is the head of } e_j, \\ -1 & \text{if } x_i \text{ is the tail of } e_j, \\ 0 & \text{otherwise.} \end{cases}$$

Let $x_i=(y_i, z_i)\in R^2$, and set

$$a_j=(a_{1j}, \cdots, a_{nj}), \quad y=(y_1, \cdots, y_n)^T, \quad z=(z_1, \cdots, z_n)^T.$$

Also let $v(x_i)=(u_i, v_i)$ be an assignment of velocities to the vertices and $u=(u_1, \cdots, u_n)$, $v=(v_1, \cdots, v_n)$. Then (1) can be written as

$$(a_jy)(a_ju)+(a_jz)(a_jv)=0, \qquad j=1, \cdots, m. \tag{3}$$

The degree of freedom of G is then the dimension of the solutions of this system of linear equations in $u_1, \cdots, u_n, v_1, \cdots, v_n$, less 3. Since the number of variables is $2n$, we see that the degree of freedom of G is $(2n-3)$-rank of (3). So the main problem is to calculate the rank of (3), i.e., the number of linearly independent $2n$-dimensional vectors of the form

$$((a_jy)a_j, (a_jz)a_j), \qquad 1\leqq j\leqq m. \tag{4}$$

This is of course easily done if y and z, i.e., the coordinates of points of G are explicitly given. But now we are interested in the case when G is generic, i.e., the entries of y and z are algebraically independent transcendentals. In this case the rank of (3) is independent of the actual choice of these transcendentals (since by the definition of algebraic dependence, a subdeterminant is 0 if and only if it is identically 0 if the coordinates of the vertices are considered as variables). So the *generic rank* of (3) depends on the graph G only (below we shall see that it depends on the polygon-matroid of G only). An (abstract) graph G will be called *stiff* if the generic structures isomorphic to G are rigid.

2. Results. The following theorem is a slight generalization of a theorem of Laman [6] (see also [2]). We give here an independent proof.

THEOREM 1. *The generic degree of freedom of a graph G with n vertices is*

$$2n-3-\min \sum_{i=1}^{k} (2|V(G_i)|-3),$$

where the minimum extends over all systems $\{G_1, \cdots, G_k\}$ of subgraphs such that $G_1\cup \cdots \cup G_k=G$.

It is easy to notice that it would suffice to extend the minimum over those systems $\{G_1, \cdots, G_k\}$ which consist of edge-disjoint spanning subgraphs.

This theorem gives a "good characterization" of the generic degree of freedom of a graph G. In fact, to prove that this number is at most f, it suffices to exhibit one particular realization of G as a (nongeneric) planar structure for which it has degree of freedom at most f; and it is not difficult to show that there exists such a realization in which the coordinates of the points are natural numbers not exceeding n (for this part we do not need the theorem). On the other hand, to prove that the generic degree of freedom is at least f, it suffices to exhibit one particular decomposition $G=G_1\cup \cdots \cup G_k$ of G into subgraphs such that

$$\sum_{i=1}^{k} (2|V(G_i)|-3)\leqq 2n-3-f.$$

The theorem guarantees that such a decomposition always exists.

But this result can also be used to obtain a polynomial-bounded algorithm to determine the generic degree of freedom of a graph. In fact, in [5] a polynomial algorithm is described which, given a submodular set-function φ defined on the subsets of a set S, and nonnegative on the nonempty subsets, computes

$$\min\left\{\sum_{i=1}^{k}\varphi(S_i)\colon S_1,\cdots,S_k \text{ is a partition of } S \text{ into nonempty subsets}\right\}.$$

One may apply this to the set-function φ defined on the subsets of $S=E(G)$ by $\varphi(Y)=2|V(Y)|-3$ (here $V(Y)$ denotes the set of vertices met by the edges in Y).

The algorithm in [5] mentioned above is rather complicated and inefficient (although polynomial). Further analysis of the result of Theorem 1 leads to a more combinatorial procedure.

A set $Y\subseteq E(G)$ is called *independent* if

$$f(Y)=2n-3-|Y|$$

(i.e., if every edge of Y takes away one degree of freedom). If G is a generic structure then the independence of a subset of edges depends on the graph structure of G only; so if we are given an abstract graph, we can define the *generic independence* of a set of its edges as the independence of this set of edges in a generic realization of the graph.

It is well known [3] that the independent subsets of edges form the independent sets of a matroid, and that

$$\begin{aligned} f(G) &= \min\{f(Y)\colon Y\subseteq E(G),\ Y \text{ independent}\} \\ &= 2n-3-\max\{|Y|\colon Y\subseteq E(G),\ Y \text{ independent}\} \end{aligned} \tag{5}$$

holds for every structure G. Thus to determine $f(G)$ it suffices to have an algorithm to check whether or not a set Y of edges is independent (using the fact that in a matroid every inclusionwise maximal independent set is maximum). Theorem 1 will imply

COROLLARY 1. *G is generic independent if and only if*

$$2|V(Y)|-3\geqq|Y| \tag{6}$$

holds for every $Y\subseteq E(G)$.

A possibility to check whether a given set Y of edges is generic independent is to find the minimum of the submodular set-function

$$\varphi(Y)=2|V(Y)|-3-|Y|$$

over $Y\subseteq X$. It is possible to find the minimum of a submodular set-function in polynomial time (see [5]). But there will be a simpler—more combinatorial—possibility to solve this problem. Observing that condition (6) is very similar to the condition which guarantees that G is decomposable into two forests (Nash–Williams [8]), we may rephrase this result as follows:

COROLLARY 2. *G is generic independent if and only if doubling any edge of G results in a graph which is the union of two forests.*

Applying an algorithm due to Edmonds [4], we can check whether a graph is the union of two forests. Running this algorithm $|E(G)|$ times we can decide whether or not G is generic independent.

The next corollary is the theorem of Laman mentioned in the introduction. It should be noted that one could derive Theorem 1 from Laman's theorem as well.

COROLLARY 3. *A graph is minimal stiff (or equivalently, both stiff and generic independent) if and only if*

(i) $|E(G)| = 2|V(G)| - 3$, *and*

(ii) $|E(H)| \leqq 2|V(H)| - 3$ *for every subgraph* H *of* G.

Finally, a fifth version of the same thing characterizes rigidity:

COROLLARY 4. *A graph* G *is stiff if and only if*

$$\sum_{i=1}^{k} (2|V(G_i)| - 3) \geqq 2|V(G)| - 3$$

holds for every system of subgraphs G_i *such that* $G_1 \cup \cdots \cup G_k = G$.

Even though the conditions given in the theorem and its corollaries are graph-theoretic and can be checked polynomially, it may be interesting to relate rigidity to other graph-theoretic properties. One such property is connectivity. This is also motivated by the trivial fact that among structures restricted to the line, connectivity and rigidity are equivalent.

THEOREM 2. *Every 6-connected graph is stiff.*

We remark here that the number 6 is the best possible:

Example. Let G_0 be a 5-regular 5-connected graph on v points. Split every vertex of G_0 into 5 vertices of degree 1, and identify these 5 vertices with the vertices of a complete 5-graph. The resulting graph G is 5-connected and not stiff. It is easy to see that G is 5-connected. To show that G is not stiff, we use Corollary 4: if $G_1, \cdots, G_v$ are the complete 5-graphs corresponding to the vertices of G_0, and $G_{v+1}, \cdots, G_k$ are the subgraphs consisting of one edge of G_0 each (clearly $k = 7v/2$), then

$$\sum_{i=1}^{k} (2|V(G_i)| - 3) = 7v + \frac{5}{2}v = \frac{19}{2}v < 2|V(G)| - 3 = 10v - 3$$

if $v > 6$. One specific example is shown in Fig. 1.

Finally, we formulate two results in combinatorial linear algebra (or in the theory of representable matroids). The first of these was proved in [7]. Let $A_1, \cdots, A_n$ be subspaces of the n-dimensional linear space F^n over a field F, and H a hyperplane (i.e., $(n-1)$-dimensional subspace) in the same linear space. We say that H is *in general*

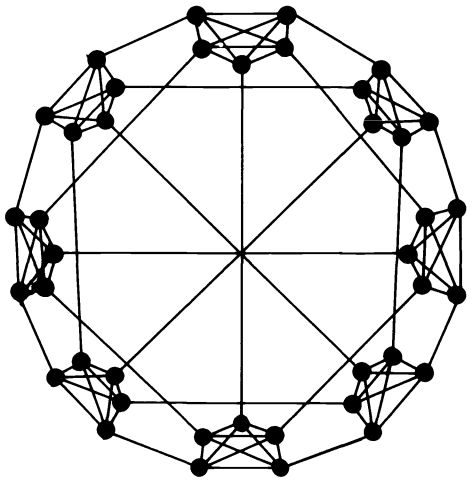


FIG. 1. *A 5-connected nonstiff graph.*

position with respect to $A_1, \cdots, A_n$ if there exists a basis for every A_i and an equation $c \cdot x = 0$ for H (i.e., c is a normal vector of H) such that the entries of c are algebraically independent over the field generated by all entries of all vectors in the bases of $A_1, \cdots, A_n$.

THEOREM 3. *Let $A_1, \cdots, A_n$ be subspaces of F^n and let H be a hyperplane in F^n in general position with respect to $A_1, \cdots, A_m$. Then*

$$\dim \langle H \cap A_1, \cdots, H \cap A_m \rangle = \min \sum_{i=1}^{k} (\dim \langle A_r : r \in N_i \rangle - 1),$$

where $\{N_1, \cdots, N_k\}$ ranges over all partitions of $\{1, \cdots, m\}$ into nonempty subsets.

This theorem will serve as a lemma to prove the following, which is an algebraic generalization of Theorem 1:

THEOREM 4. *Let F be a field, $a_1, \cdots, a_m \in F^n$, and let $y, z \in F^n$ such that the entries of y and z are algebraically independent over the field F_0 generated by the entries of $a_1, \cdots, a_m$. Then the dimension of the subspace spanned by the vectors*

$$((a_i y)a_i, (a_i z)a_i) \in F^{2n}$$

is given by

$$\min \sum_{i=1}^{k} (2 \cdot \dim \langle a_r : r \in N_i \rangle - 1),$$

where $\{N_1, \cdots, N_k\}$ ranges over all partitions of $\{1, \cdots, m\}$ into nonempty subsets.

3. Proofs.

Proof of Theorem 4. Let

$$A_i = \{(\lambda a_i, \mu a_i) : \lambda, \mu \in F\} \subseteq F^{2n}$$

and

$$H = \{(x, x') \in F^{2n} : x \cdot z = x' \cdot y\}.$$

Note that $((a_i y)a_i, (a_i z)a_i) \in A_i \cap H$, and since $\dim A_i \cap H = 1$, it follows that $A_i \cap H$ consists of the multiples of the vector $((a_i y)a_i, (a_i z)a_i)$. So

$$\dim \langle ((a_i y)a_i, (a_i z)a_i) : i = 1, \cdots, m \rangle = \dim \langle H \cap A_i : i = 1, \cdots, m \rangle.$$

Furthermore, the coefficients of the linear equation defining H are the entries of y and z, and these are by assumption algebraically independent over the field F_0 generated by the entries of the bases $\{(a_i, 0), (0, a_i)\}$ of A_i. So Theorem 3 applies and we get that

$$\begin{aligned} \dim \langle H \cap A_i : i = 1, \cdots, m \rangle &= \min \sum_{i=1}^{k} (\dim \langle A_r : r \in N_i \rangle - 1) \\ &= \min \sum_{i=1}^{k} (2 \dim \langle a_r : r \in N_i \rangle - 1), \end{aligned}$$

as claimed. □

Proof of Theorem 1. By (4), we want to determine

$$\dim \langle ((a_j y)a_j, (a_j z)a_j) : j = 1, \cdots, m \rangle.$$

By Theorem 3, this is equal to

$$\min \sum_{i=1}^{k} (2 \dim \langle a_r : r \in N_i \rangle - 1),$$

where $\{N_1, \cdots, N_k\}$ ranges over all partitions of $\{1, \cdots, m\}$ into nonempty subsets. By a fundamental fact in matroid theory, this is equal to

$$\min \sum_{i=1}^{k} (2 \cdot r\{e_j : j \in N_i\} - 1),$$

where r denotes the rank function of the polygon matroid of the graph G. But

$$\min \sum_{i=1}^{k} (2r\{e_j : j \in N_i\} - 1) = \min \sum_{i=1}^{k} (2|V(\{e_j : j \in N_i\})| - 3).$$

In fact, each term on the right-hand side is at least as large as the corresponding term on the left-hand side, and the partition $N_1, \cdots, N_k$ minimizing the left-hand side is automatically such that the edges corresponding to each N_i form a connected graph, and so the minimum term on the left-hand side is equal to the corresponding term on the right-hand side. This completes the proof. □

Proof of Corollary 1. By (5), G is generic independent if and only if its generic degree of freedom is $2n-3-m$. By Theorem 1, this is equivalent to the condition that for every system of subgraphs $G_1, \cdots, G_k$ such that $G_1 \cup \cdots \cup G_k = G$, we have

$$\sum_{i=1}^{k} (2|V(G_i)| - 3) \geqq m. \tag{7}$$

We show that (7) is equivalent to the condition given in the theorem. Assume first that (7) holds. Let H be an arbitrary subgraph of G. Choose $G_1 = H$ and let $G_2, \cdots, G_k$ be the subgraphs consisting of one edge of $E(G) - E(H)$ each. Then (7) implies

$$2|V(G_1)| - 3 + (k-1) \geqq m.$$

But clearly $k = m - |E(H)| + 1$, so this means

$$2|V(H)| - 3 \geqq |E(H)|.$$

Conversely, suppose that the condition in the theorem holds. Then

$$\sum_{i=1}^{k} (2|V(G_i)| - 3) \geqq \sum_{i=1}^{k} |E(G_i)| \geqq E(G)$$

follows for arbitrary subgraphs $G_1, \cdots, G_k$ such that $G_1 \cup \cdots \cup G_k = G$; i.e., (7) is true. □

Proof of Corollary 2. Suppose first that G is independent. Then by Corollary 1,

$$2|V(H)| - 3 \geqq |E(H)|$$

holds true for every subgraph H. If we double an edge of G, then every subgraph H of the resulting graph G' will satisfy the slightly weaker inequality

$$2|V(H)| - 2 \geqq |E(H)|.$$

By the theorem of Nash–Williams [8], this implies that G is the union of two forests.

Conversely, assume that doubling an edge results in a graph which is the union of two forests, for every edge. Let H be a subgraph. Double an edge of H and decompose the resulting graph into two forests. This yields two subforests H_1 and H_2 of H such that $H_1 \cup H_2 = H$ and H_1 and H_2 have one edge in common. Hence

$$\begin{aligned} |E(H)| = |E(H_1)| + |E(H_2)| - 1 &\leqq (|V(H_1)| - 1) + (|V(H_2)| - 1) - 1 \\ &\leqq 2|V(H)| - 3. \end{aligned}$$

Thus, by Corollary 1, G is generic independent. □

Proof of Corollary 3. If G is stiff and generic independent, then (ii) follows by Corollary 1 and (i) follows by (5) and (ii). Conversely, if (i) and (ii) hold, then G is generic independent by (ii) and Corollary 1, and it is stiff by (i) and (5). □

Proof of Corollary 4. Trivial by Theorem 1. □

Proof of Theorem 2. Suppose that G is a 6-connected nonstiff graph. We may assume that G has the least possible number n of points among all such graphs. Also we may assume that G has the largest number of edges among all such graphs on n points.

By Corollary 4, there exist subgraphs $G_1, \cdots, G_k$ of G such that $G_1 \cup \cdots \cup G_k = G$ and

$$\sum_{i=1}^{k} (2|V(G_i)|-3) < 2n-3. \tag{8}$$

We may assume that $G_1, \cdots, G_k$ are induced subgraphs. Then it follows that they must be complete subgraphs, since adding an edge spanned by $V(G_i)$ to G would result in a graph which is also 6-connected, nonstiff (since (8) still holds) and has more edges than G, which is impossible.

We show that every vertex of G occurs in at least two subgraphs G_i. Suppose indirectly that $v \in V(G)$ is a vertex of G_1 (say), but not of $G_2, \cdots, G_k$. Let $G' = G - v$, $G'_1 = G_1 - v$, $G'_2 = G_2, \cdots, G'_k = G_k$. Then $G' = G'_1 \cup \cdots \cup G'_k$, and by (8),

$$\sum_{i=1}^{k} (2|V(G'_i)|-3) < 2|V(G')|-3.$$

So G' is not stiff and so by the minimality of the number of points of G, the graph G' cannot be 6-connected, i.e., there exists a set $T \subseteq V(G')$ such that $|T| \leqq 5$ and $G' - T$ is disconnected. Let $G' - T = H_1 \cup H_2$, where H_1 and H_2 are vertex-disjoint nonempty subgraphs. Since $G - T$ is connected, v must be adjacent to at least one point v_i of H_i for both $i = 1$ and 2. But then $v_1, v_2 \in V(G_1)$ and so v_1 and v_2 must be adjacent (since G_1 is a complete graph). So T does not separate v_1 and v_2 in G', a contradiction.

Let $v \in V(G)$. Since G is 6-connected, v has degree at least 6, and so

$$\sum_{V(G_i) \ni v} (|V(G_i)|-1) \geqq 6. \tag{9}$$

Hence we deduce

$$\sum_{V(G_i) \ni v} \left(2 - \frac{3}{|V(G_i)|}\right) \geqq 2. \tag{10}$$

In fact, let (say) $v \in V(G_1), \cdots, V(G_d)$, $v \notin V(G_{d+1}), \cdots, V(G_k)$. Without loss of generality assume that $|V(G_1)| \geqq \cdots \geqq |V(G_d)|$. By the above, $d \geqq 2$. Each term in (10) is at least $\frac{1}{2}$, so if $d \geqq 4$ then (10) is obvious. If $d = 3$ then (9) implies that $|V(G_1)| \geqq 3$, and so the left-hand side of (10) is at least $1 + (\frac{1}{2}) + (\frac{1}{2}) = 2$. Finally, if $d = 2$ then (9) implies that $|V(G_1)| \geqq 4$, and also that in the cases $|V(G_1)| = 4$, 5 and $\geqq 6$ we have $|V(G_2)| \geqq 4$, 3 and 2, respectively. So the left-hand side of (10) is at least $(\frac{5}{4}) + (\frac{5}{4})$, $(\frac{7}{5}) + 1$ and $(\frac{3}{2}) + (\frac{1}{2})$, respectively. This proves (10).

Now, summing (10) for every vertex v we get

$$\sum_{i=1}^{k} |V(G_i)|\left(2 - \frac{3}{|V(G_i)|}\right) = \sum_{i=1}^{k} (2|V(G_i)|-3) \geqq 2n,$$

which contradicts (8). □

4. Concluding remarks. The proof of Theorem 2 would also yield that if we delete 3 edges from a 6-connected graph, the resulting graph is still stiff. On the other hand, deleting 4 "vertical" edges from the 6-connected graph in Fig. 2 we get a nonstiff graph.

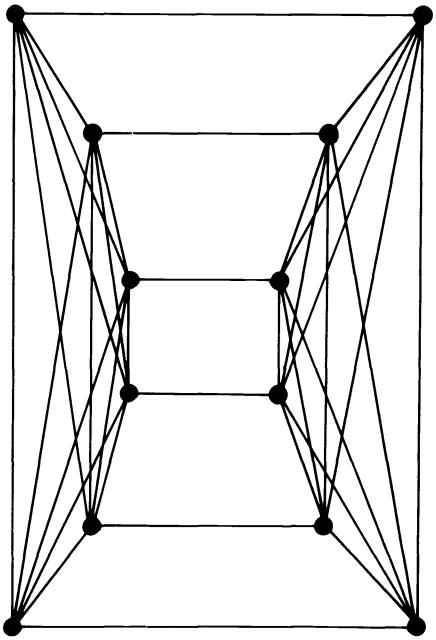


FIG. 2. *A 6-connected graph which does not remain stiff if 4 edges are deleted.*

It is natural to ask for an extension of these results to 3-dimensional structures. In spite of considerable effort on the part of several people, the problem of extending Laman's theorem to higher dimensions is still open. The proof given here, being more algebraic in nature than previously found proofs, may offer new ways of approach to the higher-dimensional case, even though the authors were unable to find these. It seems more promising to generalize Theorem 2 to the space:

Conjecture. There exists a constant c such that every c-connected graph is generic rigid in the 3-space.

Perhaps $c = 12$; more generally, perhaps $c = d(d+1)$ for the d-dimensional space (for $c = d(d+1)-1$ a counter-example similar to the one given in § 2 works).

REFERENCES

[1] L. ASIMOW AND B. ROTH, *The rigidity of graphs*, Trans. Amer. Math. Soc., 254 (1978), pp. 279–289.
[2] ———, *The rigidity of graphs* II, preprint.
[3] H. CRAPO, *Structural rigidity*, Structural Topology, 1 (1979), pp. 26–45.
[4] J. EDMONDS, *Minimum partition of a matroid into independent subsets*, J. Res. Nat. Bur. Standards Sect. B, 69 (1965), pp. 67–72.
[5] M. GRÖTSCHEL, L. LOVÁSZ AND A. SCHRIJVER, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica, 1 (1981), pp. 169–197.
[6] G. LAMAN, *On graphs and rigidity of plane skeletal structures*, J. Engrg. Math., 4 (1970), pp. 331–340.
[7] L. LOVÁSZ, *Flats in matroids and geometric graphs*, in Combinatorial Surveys, Proc. 6th British Combinatorial Conf., P. Cameron, ed., Academic Press, New York, 1977, pp. 45–89.
[8] C. ST. J. A. NASH-WILLIAMS, *Decomposition of finite graphs into forests*, J. London Math. Soc., 39 (1964), p. 12.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 1, March 1982

0196-5212/82/0301-0010 $01.00/0

UPPER AND LOWER BOUNDS ON THE COMPLEXITY OF THE MIN-CUT LINEAR ARRANGEMENT PROBLEM ON TREES*

THOMAS LENGAUER†

Abstract. The min-cut linear arrangement problem is one of several one-dimensional layout problems for undirected graphs that may be of relevance to VSLI design.

This paper gives a polynomial time algorithm that finds a min-cut linear arrangement of trees whose cost is within a factor of 2 of optimal. For complete m-ary trees a linear time algorithm is given that finds an optimum min-cut linear arrangement.

1. Introduction. The min-cut linear arrangement problem is a fundamental one-dimensional graph layout problem. The cost measure used in the min-cut linear arrangement problem models area minimization of VLSI circuits whose components are laid out in rows.

Usually the min-cut linear arrangement problem (or MINCUT problem for short) is defined as follows (see [Ga77], [GJ79]). Given an undirected graph $G = (V, E)$ with N vertices find a labeling (also called a *layout*) $\lambda : V \to \{1, \cdots, N\}$ that minimizes the following quantity (called the *width* of the layout):

$$\max_{1 \leqq i \leqq N} |\{(v, w) \in E | \lambda(v) \leqq i < \lambda(w)\}|.$$

The graph represents a circuit and the labeling a linear layout of the circuit. The active elements of the circuit, represented by the vertices of the graph, are laid out in a row, starting at the left with the vertex having the smallest label and proceeding toward the right in ascending order of the labels. The width is the maximum number of edges (wires) passing between each pair of neighboring vertices in the layout, and gives a measure for the width and if multiplied with the number of vertices in the graph also for the area of the circuit layout. Thus the MINCUT problem applies to area minimization of such layouts.

There are several approaches to LSI and VLSI design that structure the layout task by confining themselves to laying out the active elements of the circuit in rows. (See for instance [Fe76], [PDS77].) Such layout systems typically do not achieve the smallest area possible but make chip design fast and cheap. However, the heuristics that are used in such systems to place the active elements of the circuit are either simple-minded or defy analysis. A careful study of the MINCUT problem could be a first step to improve this situation.

Here we introduce the MINCUT problem in a slightly different way. We define it as a pebble game on graphs. The game is played on an undirected graph G. Pebbles are placed on the vertices of G in a certain order. All vertices start out pebble-free and end up pebbled. The order in which the pebbles are placed on the graph constitutes a *strategy*. Each placement of a pebble is a *move*. The *reversal* $\bar{S}$ of S is the strategy that performs the moves of S in reversed order. The (*edge*) *cut* after each move in S is the set of edges that connect an unpebbled vertex with a pebbled vertex. The *cost* of S, denoted by cost (S), is the maximum size of an edge cut during the strategy. Note that cost (S) = cost $(\bar{S})$. The object is to minimize the cost of S. The minimum cost for any strategy is denoted by cost (G).

* Received by the editors November 24, 1980.
† Bell Laboratories, Murray Hill, New Jersey 07974.

The correspondence between the MINCUT problem as defined in the literature and the pebble game defined here is as follows. The strategy S lists the vertices of the graph in the order of ascending labels. Thus as time progresses in the pebble game the layout is scanned from left to right. The cut after the ith move in S is the set of edges that pass between the ith and $(i+1)$st vertex in the layout. Thus the cost of S is the width of the layout.

Considering the MINCUT problem as a game has the advantage of stressing the dynamic character of the problem. However, the above definitions vary conceptually from those of different pebble games in the literature. Usually the cost measure is the number of pebbles used and the special characteristics of the game are encoded in the pebbling rules. We could have defined our pebble game in this way. However, this would have meant that we had to allow multiple pebbles on a vertex (one for each edge on the cut). The rules would therefore have been too confusing. Hence, here we made the pebbling rules trivial, and put the "semantics" of the game into the definition of the cost measure. This makes the game intuitive and easy to work with. It has also the advantage of clearly exhibiting the affinity of the pebble game studied here to another one discussed in [Le80]. The pebble game studied in [Le80] is the vertex separator version of the pebble game defined above, and can be linked to register allocation problems for nondeterministic straight-line programs, or equivalently to black-white pebbling of dags (see [Lo79], [LeTa80]). The proofs of the results about the MINCUT problem in this paper also make use of ideas from black-white pebbling arguments. This shows that cross fertilization is possible between the areas of one-dimensional graph layout and register allocation in nondeterministic straight-line programs.

Despite its relevance to printed circuit board and VLSI layout, not many results have been established about the MINCUT problem. It is known that the MINCUT problem for general undirected graphs is NP-complete ([Ga77], [GJ79]). Harper shows in [Ha64] that the n-dimensional hypercube (with $N = 2^n$ vertices) can be pebbled optimally in the MINCUT problem in $O(N \log N)$ time. In fact this can be achieved by pebbling its vertices in ascending order with respect to their respective binary labels, if the vertices are considered as corners of the n-cube. No other nontrivial subclasses of graphs have been found for which the MINCUT problem is polynomially solvable. On the other hand the MINCUT problem has also not been shown NP-complete for any nontrivial subclass of graphs. Approximation algorithms are not known either.

This paper considers the MINCUT problem on trees. Section 2 gives a lower bound for the MINCUT problem on trees. Section 3 gives an $O(N \log N)$ time algorithm for the MINCUT problem on trees that finds a strategy whose cost is within a factor of 2 of the bound given in § 2. Section 4 gives a linear time algorithm for the MINCUT problem on complete m-ary trees that finds a strategy with an optimal cost.

2. A lower bound for the MINCUT problem on trees. In this section we prove a lower bound on cost (T) where T is an undirected tree. As the example in the Appendix shows, the lower bound proved in this section is not tight for arbitrary trees, but it will turn out to be tight for complete m-ary trees (see § 4).

We start out by arbitrarily choosing one vertex r of the tree T as the root. We will call the neighbors of the root r its *children* and denote them by $r_1, \cdots, r_m$ where m is the degree of r. With $T_1, \cdots, T_m$ we will denote the rooted subtrees *induced* by $r_1, \cdots, r_m$. The tree T_i is the tree containing r_i after we delete r and all its edges from T. The root of T_i is r_i. The *height* h of T is the length of the longest path from the root to a leaf.

After thus making T into a rooted tree, we will use an argument similar to the one given in [LeTa80] to prove a lower bound on the number of pebbles needed to pebble a directed rooted tree with black-white pebbles. The lower bound will be proved by induction on the height of T. However, just inductively computing the lower bound will not be enough to carry through the induction. We will use a parity argument. In addition to computing the bound we will keep track of where the maximum cut-sizes occur in optimal pebbling strategies.

Specifically we will compute a value $L(T)$ which is about twice as great as the lower bound. Furthermore, if $L(T)$ is even, then pebbling strategies achieving the lower bound may exist for which the maximum cut occurs only *after* the root is pebbled. If $L(T)$ is odd, then all pebbling strategies for T that achieve the lower bound are such that the maximum cut occurs both *before and after* the root is pebbled. (Note that the tree consisting of just one vertex with no edges is in the following called the *trivial* tree.)

DEFINITION 1. Let T be an undirected rooted tree with root r. Let the two quantities $L(T)$ and $\alpha(T)$ be defined as follows:

$$L(T) = -1 \quad \text{and } \alpha(T) = 1 \quad \text{if } T \text{ is the trivial tree.}$$

Otherwise let the root r of T have degree $m \geqq 1$. Let its children r_i be ordered such that $L(T_1) \geqq \cdots \geqq (T_m)$, where T_i is the subtree of T induced by r_i. Then

$$L(T) = \max\{L(T_i) + i - 1 + \alpha(T_i) | 1 \leqq i \leqq m\},$$

$$\alpha(T) = \begin{cases} 1 & \text{if } m \text{ is even and } L(T) = m - 1, \\ 0 & \text{otherwise.} \end{cases}$$

The quantity $\alpha(T)$ encodes a special situation which occurs if T is "shallow" and the degree of r is even. In this case optimal pebbling strategies will essentially pebble half of the children of the root before the root and the other half after the root. The maximum cut-sizes will occur both directly before and directly after the root is pebbled. $\alpha(T)$ also takes care of a technical adjustment that has to be made if T is the trivial tree.

The quantity $\alpha(T)$ is needed for tightly matching the upper bound on the linear arrangement of complete m-ary trees given in § 4. Note that $\alpha(T)$ is not necessary for proving the result given in § 3.

The following theorem is the heart of the lower bound proof and justifies the above definitions. (For an example see the Appendix.)

THEOREM 2. *Let T be a rooted undirected tree. Consider any strategy that pebbles T. Let t be the time at which the root r is pebbled.*

(a) *If $\alpha(T) = 1$ then at $t-1$ or at t at least $\lfloor L(T)/2 \rfloor + 2$ edges are on the cut, or else both at $t-1$ and at t, $\lfloor L(T)/2 \rfloor + 1$ edges are on the cut.*

(b) *If during $[0, t-1]$ at most $\lfloor L(T)/2 \rfloor$ edges are always on the cut then sometime in $[t, \infty)$ at least $\lceil L(T)/2 \rceil + 1$ edges are on the cut.*

Note that statement (b) in this theorem is more stringent when $L(T)$ is odd than when $L(T)$ is even. When $L(T)$ is even the theorem merely says that if $\lfloor L(T)/2 \rfloor + 1$ edges did not appear on the cut before t then they must appear on the cut sometime after t. When $L(T)$ is odd the theorem says that any strategy with a maximum cut of size $\lceil L(T)/2 \rceil + 1$ must have a maximum cut once before and once after t.

Proof. By induction on the height h of T. Trivial for $h = 0$.

Let $h \geqq 1$. Then $m \geqq 1$. Statement (a) is trivially true since if $\alpha(T) = 1$ then $\lfloor L(T)/2 \rfloor + 1$ is exactly one half times the number of edges leaving r. For the proof

of (b), let j be the smallest value such that $L(T) = L(T_j) + j - 1 + \alpha(T_j)$. Deleting subtrees $T_{j+1}, \cdots, T_m$ from T neither changes $L(T)$, nor does it increase the cut. Thus we can, without loss of generality, assume that $j = m$ and $L(T_i) \geqq L(T) - m + 1 - \alpha(T_i)$ for $1 \leqq i \leqq m$.

We reorder the children of r in the following fashion. We select a critical value c such that $1 \leqq c \leqq m$, depending on the case we want to prove. Then we let $T_1, \cdots, T_{c-1}$ be the $c-1$ subtrees that received a pebble first. The remaining $m-c+1$ subtrees $T_c, \cdots, T_m$ are numbered in the sequence in which they receive their last pebble. The first of these subtrees, T_c, is called the critical subtree. We will find times at which there are many edges on the cut in T by finding times at which there are many edges on the cut in T_c.

We have to make a case distinction:

Case 1. $L(T)$ is odd. Choose $c = \lceil m/2 \rceil$. Assume that at most $\lfloor L(T)/2 \rfloor$ edges are on the cut before t. Since $c-1$ subtrees receive their first pebble before T_c, T_c has during $[0, t-1]$ a cut-size of at most $\lfloor L(T)/2 \rfloor - c + 1$.

Case 1.1. $\alpha(T_c) = 0$. In this case

$$\begin{aligned} \lfloor L(T)/2 \rfloor - c + 1 &= (L(T)-1)/2 + \lfloor -m/2 \rfloor + 1 \\ &= \lfloor (L(T)-1)/2 - (m/2-1) \rfloor \\ &= \lfloor (L(T)-m+1)/2 \rfloor \leqq \lfloor L(T_c)/2 \rfloor. \end{aligned} \tag{*}$$

Case 1.1.1. Assume that before r_c is pebbled at most $\lfloor L(T_c)/2 \rfloor$ edges are on the cut inside T_c. Then at some time t' after r_c is pebbled at least $\lceil L(T_c)/2 \rceil + 1$ edges are on the cut in T_c by induction. Furthermore $t' > t$ holds because of (*). Thus in addition, for each of the trees T_i, $(i = c+1, \cdots, m)$ at least one edge is on the cut at time t'. This is either an edge inside T_i (if T_i has received a pebble before t') or the edge between r and r_i (since r has a pebble at t'). Thus the cut at t' has size at least $\lceil L(T_c)/2 \rceil + 1 + m - c$.

Case 1.1.2. Otherwise. At some time $t' > t$ before r_c is pebbled there at least $\lfloor L(T_c)/2 \rfloor + 1$ edges on the cut in T_c. In addition, at t' as in Case 1.1.1, there is one edge on the cut in T for each T_i $(i = c+1, \cdots, m)$. Furthermore the edge between r and r_c is on the cut at t'. Thus the cut at t' has size at least

$$\lfloor L(T_c)/2 \rfloor + 2 + m - c \geqq \lceil L(T_c)/2 \rceil + 1 + m - c.$$

In both cases the number of edges on the cut at t' is at least

$$\begin{aligned} \lceil L(T_c)/2 \rceil + 1 + m - c &= \lceil L(T_c)/2 \rceil + 1 - \lceil -m/2 \rceil \\ &\geqq \lceil (L(T)-m+1)/2 \rceil - \lceil -m/2 \rceil + 1 \\ &= (L(T)+1)/2 + 1 = \lceil L(T)/2 \rceil + 1, \end{aligned}$$

and the theorem is proved for Case 1.1.

Case 1.2. $\alpha(T_c) = 1$. At most

$$\lfloor L(T)/2 \rfloor - c + 1 = \lfloor (L(T)-m+1)/2 \rfloor \leqq \lfloor L(T_c)/2 \rfloor + 1$$

edges are on the cut inside T_c before t. Furthermore the root of T_c cannot have been pebbled before t, because in that case either directly before r_c is pebbled $\lfloor L(T_c)/2 \rfloor + 2$ edges are on the cut in T_c, or directly after r_c is pebbled $\lfloor L(T_c)/2 \rfloor + 1$ edges are on the cut in T_c and the edge between r and r_i is on the cut. In both cases the total

cut-size around the time that r_c is pebbled would be at least (observe that $L(T_c)$ is odd)

$$\begin{aligned}\lfloor L(T_c)/2\rfloor + c + 1 &= (L(T_c)-1)/2 + \lfloor (m+1)/2\rfloor + 1\\ &= \lfloor (L(T_c)+m)/2\rfloor + 1 \geqq \lfloor L(T)/2\rfloor + 1.\end{aligned}$$

Thus r_c is pebbled after t. Let $t' > t$ be the time when r_c is pebbled. As above at time $t'-1$ or time t' we have a cut of size at least

$$\lfloor L(T_c)/2\rfloor + 2 + m - c,$$

where one edge is contributed as in Case 1.1 by each of the trees t_i $(i = c+1, \cdots, m)$. Now

$$\begin{aligned}\lfloor L(T_c)/2\rfloor + 2 + m - c &\geqq \lfloor (L(T)-m)/2\rfloor + 2 + \lfloor m/2\rfloor\\ &= (L/T)+1)/2 + \lfloor -(m+1)/2\rfloor + 2 + \lfloor m/2\rfloor\\ &= \lceil L(T)/2\rceil - \lceil (m+1)/2\rceil + 2 + \lfloor m/2\rfloor\\ &= \lceil L(T)/2\rceil + 1.\end{aligned}$$

This completes Case 1.

Case 2. $\lfloor L(T)\rfloor$ is even. Choose $c = \lfloor m/2\rfloor + 1$. Assume that at most $\lfloor L(T)/2\rfloor$ edges are on the cut before t. As in Case 1, T_c has never before t more than $\lfloor L(T)/2\rfloor - c + 1$ edges on the cut.

Case 2.1. $\alpha(T_c) = 0$. In this case

$$\begin{aligned}\lfloor L(T)/2\rfloor - c + 1 &= L(T)/2 - \lceil (m-1)/2\rceil\\ &= L(T)/2 + \lfloor -(m-1)/2\rfloor\\ &= \lfloor (L/T) - m + 1)/2\rfloor\\ &\leqq \lfloor L(T_c)/2\rfloor.\end{aligned}$$

As in Case 1.1 we get a case distinction. In both cases the cut at t' is at least of size

$$\begin{aligned}\lceil L(T_c)/2\rceil + 1 + m - c &= \lceil L(T_c)/2\rceil + \lceil m/2\rceil\\ &\geqq \lceil (L(T) - m + 1)/2\rceil + \lceil m/2\rceil\\ &= L(T)/2 + \lceil -(m-1)/2\rceil + \lceil m/2\rceil\\ &= L(T)/2 - \lfloor (m-1)/2\rfloor + \lceil m/2\rceil\\ &= \lceil L(T)/2\rceil + 1.\end{aligned}$$

Case 2.2. $\alpha(T_c) = 1$. As in Case 1.2 at most $\lfloor L(T_c)/2\rfloor + 1$ edges are on the cut inside T_c before t. Furthermore, as in Case 1.2, if r_c were pebbled before t, then either directly before or directly after r_c is pebbled the total cut would have size at least

$$\begin{aligned}\lfloor L(T_c)/2\rfloor + c + 1 &= (L(T_c)-1)/2 + \lfloor m/2\rfloor + 2\\ &= \lfloor (L(T_c)+m-1)/2\rfloor \geqq \lfloor L(T)/2\rfloor + 1.\end{aligned}$$

Thus r_c is pebbled after t. Let $t' > t$ be the time that r_c is pebbled. As in Case 1.2 at time $t'-1$ or at time t' the cut has size at least

$$\begin{aligned}\lfloor L(T_c)/2\rfloor + 2 + m - c &\geqq \lfloor (L(T)-m)/2\rfloor + 1 + \lceil m/2\rceil\\ &= L(T)/2 + \lfloor -m/2\rfloor + 1 + \lceil m/2\rceil \leqq \lceil L(T)/2\rceil + 1.\end{aligned}$$

This completes the proof of the theorem. □

The parity argument is a necessary enhancement of the induction hypothesis for the induction to succeed. It can be left out in the statement of the lower bound, however.

COROLLARY 3. *For an undirected rooted tree T we have*

$$\text{cost}(T) \geqq \left\lfloor \frac{L(T)}{2} \right\rfloor + 1.$$

Since we are using recursion on T, it is often of value to include the edge between r_i and r when we consider T_i. If T_i is pebbled completely before r is pebbled, this edge is on the cut starting from the time that r_i is pebbled until the pebbling of T_i is completed. In this case, when we confine ourselves to considering T_i we can think of its root r_i being permanently *anchored* with an edge to an additional unpebbled vertex. This observation motivates the following definition.

DEFINITION 4. We call an undirected rooted tree *anchored* if we consider its root to be permanently attached to an unpebbled vertex in the MINCUT problem. The edge attaching the root to the unpebbled vertex is called an *anchor*. We denote the corresponding min-cut by $\text{cost}^a(T)$.

Note that in terms of the layout, a tree is anchored if its root is considered to have an additional long edge attached that reaches to the right across the whole layout.

COROLLARY 5. *For any undirected rooted tree T we have*

$$\text{cost}^a(T) \geqq \left\lceil \frac{L(T)}{2} \right\rceil + 1.$$

Proof. Follows from the note preceding the proof of Theorem 2, and from the fact that for pebbling strategies in the MINCUT problem $\text{cost}(S) = \text{cost}(\bar{S})$. □

3. An approximation algorithm for the MINCUT problem on trees. This section introduces a recursive algorithm SHUFFLE-PEBBLE that pebbles an undirected rooted (anchored or unanchored) tree T in the MINCUT problem and computes the cost of the strategy. SHUFFLE-PEBBLE pebbles T by computing the strategies for the subtrees induced by the children of r. Then the subtrees are ordered in decreasing difficulty and the pebbling strategies are concatenated as follows. First all odd-numbered subtrees are pebbled in order of decreasing difficulty. Then r is pebbled. Then all even-numbered subtrees are pebbled in order of increasing difficulty. This idea of "shuffling" has widely been used in similar problems (see [Me78] and [Sh79]).

SHUFFLE-PEBBLE uses the following variables. It takes as input the tree T and a logical variable *anchored* that is 1 if the tree is anchored and 0 if the tree is unanchored. SHUFFLE-PEBBLE returns a permutation S of the vertices of T that represents the strategy computed, and an integer U containing the cost of the strategy. We now give a high level version of SHUFFLE-PEBBLE.

```
      proc SHUFFLE-PEBBLE (tree T; bit anchored; vertex sequence S; integer
      U):
      begin
1.      if degree(r) = 0 then
          begin
            S := r;
            U := if anchored then 1 else 0
          end
        else
2.        begin
            foreach r_i among the children of r do
              SHUFFLE-PEBBLE (T_i, 1, S_i, U_i);
```

```
3.          sort r_i such that U_1 ≧ ··· ≧ U_m;
4.          S := ∅;
            U := U' := 0;
5.          for i from 1 by 2 to 2⌊m/2⌋ − 1 do
            begin  S := S‖S_i;
              U' := max{U', U_i + (i − 1)/2}
            end;
6.          S := S‖r;
7.          for i from 2⌈m/2⌉ by −2 to 2 do
            begin  S := S‖S̄_i;
              U := max{U, U_i + i/2 − 1}
            end;
            if anchored then
              case (U' < U): S := S̄,
8.                 (U' = U): U := U + 1,
9.                 (U' > U): U := U'
              esac
10.         else   U := mac{U, U'}
          end
      end
```

It is clear that the algorithm SHUFFLE-PEBBLE constructs a valid pebbling strategy. It runs in time $O(N \log N)$ since everything except the sorting runs in linear time. We have to prove that it computes the correct cost of the strategy.

THEOREM 6. *After the call SHUFFLE-PEBBLE* $(T, anchored, S, U)$, U *is set to the cost of the strategy* S.

Proof. The theorem is trivial in the case that T is the trivial tree. If T is nontrivial, U (resp. U') maximize correctly the cut-size before (resp. after) r is pebbled in S. (Remember that the root of an anchored tree is adjacent to a permanently *unpebbled* vertex. Therefore the strategies S_i have to be reversed in statement 7.) If T is unanchored then statement 11 correctly computes U. If T' is anchored and $U' < U$ then by reversing the strategy S we can account for the anchor without increasing U. If $U' = U$ the anchor has to be accounted for separately. □

Note that SHUFFLE-PEBBLE computes $U = U_{SP}(T)$ (resp. $U = U^a_{SP}(T)$ if T is anchored) according to the following recursive scheme:

$$\left.\begin{matrix} U_{SP}(T) = 0 \\ U^a_{SP}(T) = 1 \end{matrix}\right\} \quad \text{if } T \text{ is trivial.}$$

Otherwise let the children of the root r of T be ordered such that $U^a_{SP}(T_1) \geqq \cdots \geqq U^a_{SP}(T_m)$, then

$$U_{SP}(T) = \max\left\{ U^a_{SP}(T_i) + \left\lfloor \frac{i-1}{2} \right\rfloor \middle| 1 \leqq i \leqq m \right\},$$

$$U^a_{SP}(T) = U_{SP}(T) + \beta(T), \quad \text{where}$$

$$\beta(t) = \begin{cases} 1 & \text{if the maximum for } U_{SP}(T) \text{ is achieved both with an odd and an even } i, \\ 0 & \text{otherwise.} \end{cases}$$

The following lemma uses this characterization of $U_{SP}(T)$ to show that SHUFFLE-PEBBLE comes within a factor of 2 of the optimum arrangement, and indeed within a factor 2 of the lower bound from Theorem 2.

LEMMA 7. *Let T be an unanchored tree and T^a be the anchored version of T. Then*

(a) $$U^a_{SP}(T) \leqq L(T)+2,$$

(b) $$U_{SP}(T) \leqq L(T)+1.$$

Proof. By induction on the height h of T. Trivial for $h=0$. If $h>0$ we can inductively assume the lemma to be true for all trees of height less than h.

(a) Let π be a permutation of $\{1, \cdots, m\}$ such that if $L(T_1) \geqq \cdots \geqq L(T_m)$ then $U^a_{SP}(T_{\pi(1)}) \geqq \cdots \geqq U^a_{SP}(T_{\pi(m)})$. Let the maximum for $U^a_{SP}(T)$ be achieved at i, i.e.,

(*) $$U^a_{SP}(T) = U^a_{SP}(T_{\pi(i)}) + \left\lfloor \frac{i-1}{2} \right\rfloor + \beta(T).$$

Furthermore let this i be even if possible. For $1 \leqq j \leqq i$ we have $U^a_{SP}(T_{\pi(j)}) \geqq U^a_{SP}(T_{\pi(i)})$ and by the induction hypothesis

(**) $$L(T_{\pi(j)})+2 \geqq U^a_{SP}(T_{\pi(i)}).$$

Equation (**) therefore holds for the i largest values of $L(T_j)$ and especially for $L(T_i)$; i.e.,

$$L(T_i)+2 \geqq U^a_{SP}(T_{\pi(i)}).$$

Substituting this into (*) we get

$$U^a_{SP}(T) \leqq L(T_i) + \left\lfloor \frac{i-1}{2} \right\rfloor + \beta(T) + 2.$$

If $\beta(T)=0$ then, since $\lfloor (i-1)/2 \rfloor \leqq i-1$, we have

$$U^a_{SP}(T) \leqq L(T_i) + i - 1 + 2 \leqq L(T) + 2.$$

If $\beta(T)=1$ then by the choice of i and the definition of $\beta(T)$, i is even, such that $\lfloor (i-1)/2 \rfloor \leqq i-2$ and again

$$U^a_{SP}(T) \leqq L(T_i) + i - 2 + 1 + 2 \leqq L(T_i) + 2.$$

(b) An argument analogous to that in (a) applies. We define a permutation π' such that if $L(T_1) \geqq \cdots \geqq L(T_m)$ then $U_{SP}(T_{\pi'(1)}) \geqq \cdots \geqq U_{SP}(T_{\pi'(m)})$. The case $\beta(T)=1$ above does not apply in this case. □

COROLLARY 8. SHUFFLE-PEBBLE *finds a strategy that is within a factor of* 2 *of the lower bound given in Theorem* 2.

The above analysis, even though it is rather loose on shallow trees, is almost tight on complete binary trees. For complete binary trees of T^2_h of height $h \geqq 2$ we have

$$L(T^2_h) = U^a_{SP}(T^2_h) = U_{SP}(T^2_h) + 1 = h + 1.$$

So the analysis is tight up to an additive constant of 2 which enters because of the difference between L and U_{CP} on trees of heights less than 2.

4. The MINCUT problem on complete m-ary trees. This section contains a linear algorithm OPTIMUM-PEBBLE that finds an optimum strategy for pebbling the complete m-ary tree T^m_h of height h in the MINCUT problem. The algorithm is derived from the algorithm S in [Lo79] for pebbling complete m-ary trees with black-white pebbles.

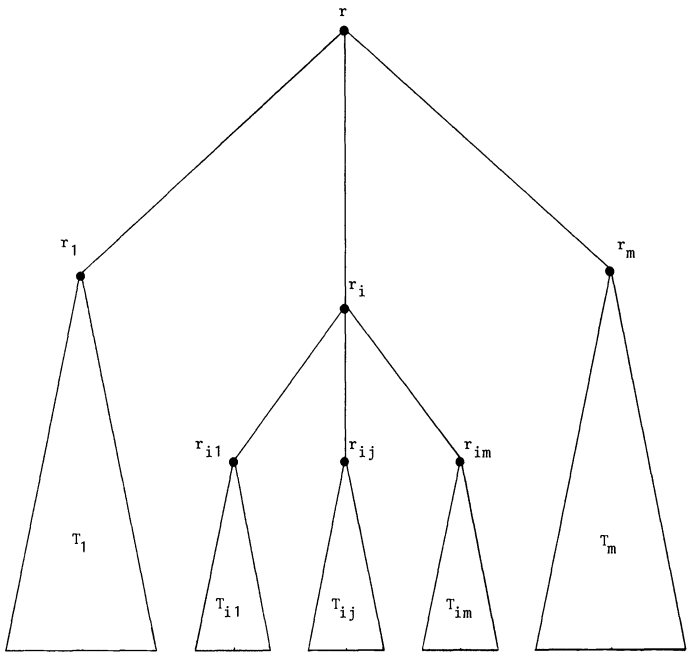


FIG. 1. *A complete m-ary tree.*

Again we consider the tree to be rooted. We will adopt the following notation in the algorithm. The root of T is denoted by r. The vertices $r_1, \cdots, r_m$ are the children of r in T_h^m. $T_1, \cdots, T_m$ are the subtrees induced by $r_1, \cdots, r_m$. For $i = 1, \cdots, m$ and $j = 1, \cdots, m$, r_{ij} are the children of r_i, T_{ij} are the corresponding subtrees (see Fig. 1).

We use variables similar to those in § 3. This time the integers m, h specify the tree; the algorithm will construct the tree T_h^m from m and h and store it in the data structure T. The logical variable *anchored* again specifies whether the tree is anchored or not. The vertex sequence S contains the computed strategy, and $U = U_{\mathrm{OP}}(T_h^m)$ (resp. $U = U_{\mathrm{OP}}^a(T_h^m)$ if T_h^m is anchored) is the cost of the computed strategy. Beside the algorithm we list as a comment the maximum cut-sizes during critical statements. Since all subtrees on a level in T_h^m are isomorphic, this time we only need one recursive call to OPTIMUM-PEBBLE. The resulting strategy can be tailored to a specific incarnation of the subtree by translation. If S' is the strategy that has been recursively computed for T_{h-2}^m, say, then we use the notation "S' on T_{ij}" to denote the translation of the strategy S' into the specific tree T_{ij}. This translation can be done in linear time in the size of T_{ij} and therefore the algorithm runs in linear time.

```
    proc OPTIMUM-PEBBLE (integer m, h; bit anchored; vertex sequence S;
    integer U):
    tree T;
1.  T := Construct m-ary tree of height h;                    cut-size
2.  if h = 0 then
      begin
        S := r;
        U := if anchored then 1 else 0
      end;
```

3. **if** $h = 1$ **then**
   begin
   $S := \emptyset$;
   for i **from** 1 **to** $\left\lceil \frac{m}{2} \right\rceil$ **do**
   $S := S \| r_i$;
   $S := S \| r$;
   for i **from** $\left\lceil \frac{m}{2} \right\rceil + 1$ **to** m **do**
   $S := S \| r_i$;
   $U :=$ **if** *anchored* **then** $\left\lfloor \frac{m}{2} \right\rfloor + 1$ **else** $\left\lceil \frac{m}{2} \right\rceil$
   end;
4. **if** $h \geqq 2$ **then**
   if *anchored* **then**
   begin
5. $S := \emptyset$;
6. OPTIMUM-PEBBLE $(m, h-2, 1, S', U')$;
7. **for** i **from** 1 **to** $\left\lceil \frac{m}{2} \right\rceil$ **do** $U' + m - 1$
   begin
   for j **from** 1 **to** $\left\lfloor \frac{m}{2} \right\rfloor$ **do**
   $S := S \| S'$ on T_{ij};
   $S := S \| r_i$;
   for j **from** $\left\lfloor \frac{m}{2} \right\rfloor + 1$ **to** m **do**
   $S := S \| \overline{S'}$ on T_{ij}
   end;
8. **for** j **from** 1 **to** $\left\lfloor \frac{m}{2} \right\rfloor$ **do** $U' + m - 1$
   $S = S \| S'$ on $T_{\lceil m/2 \rceil + 1, j}$;
9. $S := S \| r$; $2\left\lfloor \frac{m}{2} \right\rfloor + 1$
10. $S := S \| r_{\lceil m/2 \rceil + 1}$; m
11. **for** j **from** $\left\lfloor \frac{m}{2} \right\rfloor + 1$ **to** m **do** $U' + m - 1$
    $S := S \| \overline{S'}$ on $T_{\lceil m/2 \rceil + 1, j}$;
12. **for** i **from** $\left\lceil \frac{m}{2} \right\rceil + 2$ **to** m **do** $U' + m - 1$
    begin
    for j **from** 1 **to** $\left\lfloor \frac{m}{2} \right\rfloor$ **do**
    $S := S \| S'$ on T_{ij};
    $S := S \| r_i$;
    for j **from** $\left\lfloor \frac{m}{2} \right\rfloor + 1$ **to** m **do**
    $S := S \| \overline{S'}$ on T_{ij}
    end;

13. $U := \max\left\{U' + m - 1, 2\left\lfloor\frac{m}{2}\right\rfloor + 1\right\}$

end

else "(T is not anchored)"

begin

14. OPTIMUM-PEBBLE (m, $h-1$, 1, S', U');

15. **for** i **from** 1 to $\left\lceil\frac{m}{2}\right\rceil$ **do** $U' + \left\lceil\frac{m}{2}\right\rceil - 1$

$S := S \| S'$ on T_i;

16. $S := S \| r$; $\left\lceil\frac{m}{2}\right\rceil$

17. **for** i **from** $\left\lceil\frac{m}{2}\right\rceil + 1$ to m **do** $U' + \left\lceil\frac{m}{2}\right\rceil - 1$

$S := S + \overline{S'}$ on T_i;

18. $U := U' + \left\lceil\frac{m}{2}\right\rceil - 1$

end

It is clear that OPTIMUM-PEBBLE correctly computes the cost U of S. We get Table 1 for $U = U_{\mathrm{OP}}(T)$ (resp. $U = U^a_{\mathrm{OP}}(T)$).

TABLE 1

	$U^a_{\mathrm{OP}}(T^m_h)$	$U_{\mathrm{OP}}(T^m_h)$
$h=0$	1	0
$h=1$	$\left\lfloor\frac{m}{2}\right\rfloor+1$	$\left\lceil\frac{m}{2}\right\rceil$
$h=2$	$2\left\lfloor\frac{m}{2}\right\rfloor+1$	m
$h\geqq 3$	$U^a_{\mathrm{OP}}(T^m_{h-2})+m-1$	$U^a_{\mathrm{OP}}(T^m_{h-1})+\left\lceil\frac{m}{2}\right\rceil-1$

Inductively it is easy to see that for $h \geqq 2$

$$U^a_{\mathrm{OP}}(T^m_h) = \left\lceil\frac{(h-1)(m-1)}{2}\right\rceil + \left\lfloor\frac{m}{2}\right\rfloor + 1,$$

$$U_{\mathrm{OP}}(T^m_h) = \left\lceil\frac{h(m-1)}{2}\right\rceil + 1$$

Note that we have the following values for $L(T^m_h)$.

$$L(T^m_0) = 1,$$

$$L(T^m_1) = m-1,$$

$$L(T^m_h) = (h-1)(m-1) + 2\left\lfloor\frac{m}{2}\right\rfloor \quad \text{for } h \geqq 2.$$

Using these values for $L(T^m_h)$ one can easily check that the lower bounds derived in § 2 coincide with the upper bounds derived in this section. Thus OPTIMUM-PEBBLE finds optimum strategies for complete m-ary trees.

Let us compare the performance of OPTIMUM-PEBBLE and SHUFFLE-PEBBLE on complete m-ary trees. The costs of the layouts computed by SHUFFLE-PEBBLE on complete m-ary trees are as follows:

$$U^a_{\mathrm{OP}}(T^m_h) = \left\lfloor \frac{m}{2} \right\rfloor h + 1,$$

$$U_{\mathrm{OP}}(T^m_h) = \left\lfloor \frac{m}{2} \right\rfloor (h-1) + \left\lceil \frac{m}{2} \right\rceil.$$

Thus, as m increases, the ratio between the costs of the layouts computed by SHUFFLE-PEBBLE and OPTIMUM-PEBBLE decreases. Specifically we have for even m and $h \geqq 2$

$$U^a_{\mathrm{SP}}(T^m_h) = U^a_{\mathrm{OP}}(T^m_h) + \left\lfloor \frac{h}{2} \right\rfloor - 1,$$

$$U_{\mathrm{SP}}(T^m_h) = U_{\mathrm{OP}}(T^m_h) + \left\lfloor \frac{h}{2} \right\rfloor - 1.$$

If m is odd, then SHUFFLE-PEBBLE is in fact optimal on complete m-ary trees. This is not surprising: The main difference between SHUFFLE-PEBBLE and OPTIMUM-PEBBLE is that OPTIMUM-PEBBLE is careful about pebbling the root at just the right time, in order to save a little of the layout cost. However, if m is odd, the cost of the layout of T^m_h is relatively insensitive to the exact time that the root is pebbled. In particular, in both algorithms the size of the cut does not change at the time that the root is pebbled.

5. Conclusions. In § 2 of this paper we proved a lower bound for the MINCUT problem on undirected trees. In § 3 we described a polynomial time algorithm that comes within a factor of two of the lower bound on arbitrary undirected trees. This algorithm has an interesting additional property. If we select two edges of T then either both or none of the end vertices of one edge is pebbled between the end vertices of the other edge. Thus in this sense no edges "cross" over in the layout. In § 4 we gave a linear time algorithm that achieves the lower bound on complete m-ary trees. This algorithm does not have the additional property described above.

The example in the Appendix shows that neither the upper bound nor the lower bound proved in this paper about the MINCUT problem is tight in all cases. We need new insights to improve both bounds. A judicious choice of the root may be a step in the right direction, but it alone is not sufficient (see Appendix).

Approaches similar to those which are successful in the sum linear arrangement problem on trees (see [Sh79], [Ch80]) may be transferable to the MINCUT problem. In the sum linear arrangement the cost of a layout is computed by adding up the lengths of all edges. The MINCUT problem is harder to analyze than the sum linear arrangement problem. This is because the cost of the sum linear arrangement consists of contributions for each edge, where each edge contributes the distance of its end vertices in the arrangement. Therefore the modification of a layout implies a change in the cost, which can easily be estimated. In the sum linear arrangement it is in general not advantageous to move adjacent vertices in the graph far away from each other in the layout. The same is not true for the MINCUT problem. Here the modification of a layout implies a change in the cost which is harder to predict. In particular it may be possible to improve on the cost of the layout by widely separating adjacent vertices in the graph.

Appendix. As an illustration to the above discussion let us consider the following example. Let T be the undirected tree depicted in Fig. 2.

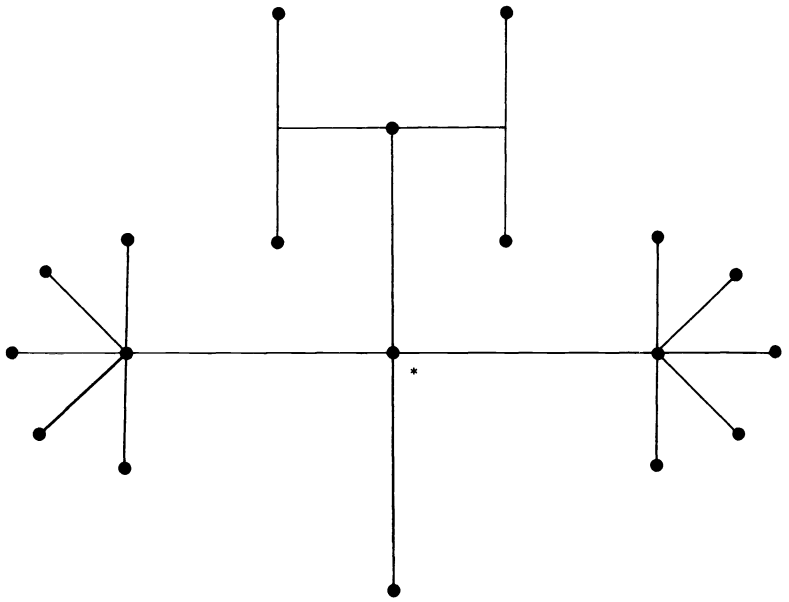


FIG. 2. *The tree T.*

If we root T at the vertex labeled with * we get the L-values given in Fig. 3. (The number next to a vertex in Fig. 3 is the L-value of the subtree induced by this vertex. The number next to an edge is the contribution of the subtree at the lower end of the edge toward the L-value of the vertex at the upper end of the edge.)
Thus if T is rooted at r then $L(T) = 5$, i.e., cost $(T) \geqq \lfloor 5/2 \rfloor + 1 = 3$.

Note that this is not a tight lower bound. In fact, no pebbling strategy with a maximum cut of size less than 4 exists. This can be seen as follows. Assume the existence of a pebbling strategy S with a maximum cut of size 3. For the subtrees T_1,

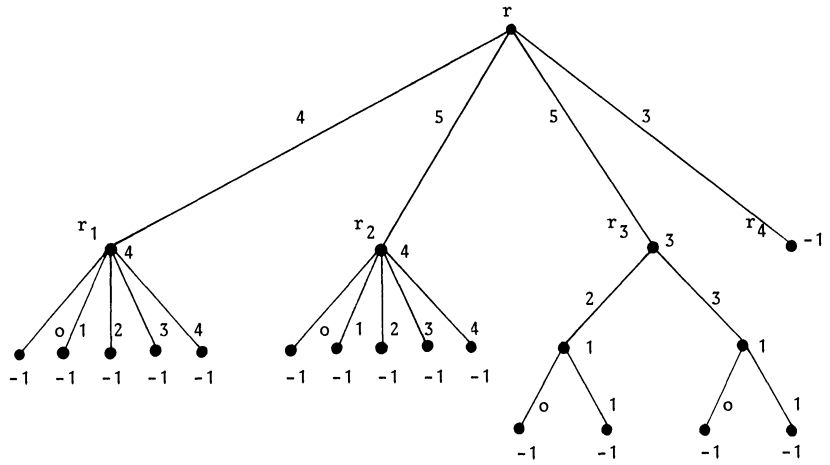


FIG. 3. *The L-values after rooting T.*

T_2, T_3 we have $\text{cost}^a(T_i) \geqq 3$, by Corollary 5. Therefore each of T_1, T_2, T_3 has to contain either the first or the last vertex pebbled in S. This is impossible, since the subtrees are disjoint, and there are only two such vertices.

Furthermore, no matter which vertex we choose as the root, $L(T)$ never exceeds the value 5. Thus the above lower bound is not tight for T.

The algorithm SHUFFLE-PEBBLE pebbles T, if rooted at the vertex labeled with *, with 4 pebbles. The pebbling strategy computed by SHUFFLE-PEBBLE corresponds (as described in the Introduction) to the layout of T depicted in Fig. 4.

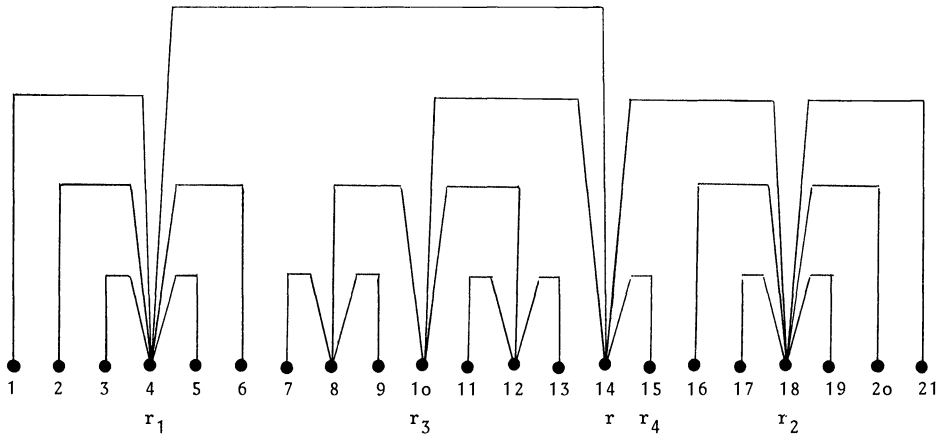


FIG. 4. *The layout computed by* SHUFFLE-PEBBLE.

The number under each vertex v is its label $\lambda(v)$, as defined in the Introduction. Note that the maximum cut-size of 4 is only achieved once, namely between the vertices labeled 11 and 12. Thus, if $L(T)$ where one greater the lower bound would be tight. In this particular example, SHUFFLE-PEBBLE finds an optimum strategy, whereas the lower bound is not tight. There are other examples, for instance complete m-ary trees, for which the lower bound is tight, whereas SHUFFLE-PEBBLE does not find optimum strategies.

Finally, in Fig. 5, we give an example of a layout computed by OPTIMUM-PEBBLE, namely the layout of the complete binary tree of height 4.

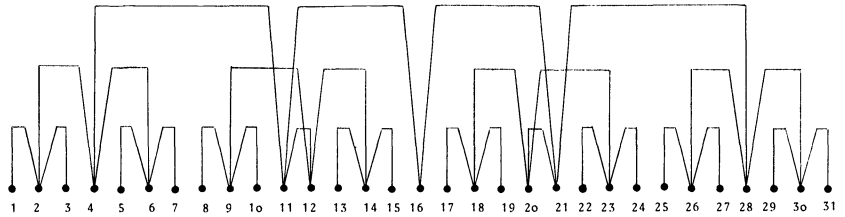


FIG. 5. *The layout of* T_4^2 *computed by* OPTIMUM-PEBBLE.

Acknowledgment. I am grateful to A. V. Aho for his many helpful suggestions that led to improvements in this presentation.

REFERENCES

[Ch 80] F. R. K. CHUNG, *On linear arrangements of trees*, Tech. Rep., Bell Laboratories, Murray Hill, NJ, 1980.

[Fe 76] A. FELLER, *Automatic layout of low-cost quick-turnaround random-logic custom LSI devices*, in Proc. 13th Design Automation Conf., San Francisco, 1976, pp. 79–85.

[Ga 77] F. GAVRIL, *Some NP-complete problems on graphs*, in Proc. 11th Conf. on Information Sciences and Systems, Johns Hopkins University, Baltimore, MD, 1977, pp. 91–95.

[GJ 79] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability, A Guide to the Theory of NP-Completeness*, Freeman, San Francisco, 1979.

[Ha 64] L. H. HARPER, *Optimal assignments of numbers to vertices*, J. Soc. Indust. Appl. Math. 12 (1964), pp. 131–135.

[Le 80] T. LENGAUER, *Relationships between pebble games on directed and indirected graphs*, Tech. Rep., Bell Laboratories, Murray Hill, NJ, 1980, Acta Informatica, to appear.

[Lo 79] M. C. LOUI, *The space complexity of two pebble games on trees*, LCS Rep. 133, MIT, Cambridge, MA, 1979.

[LeTa 80] T. LENGAUER AND R. E. TARJAN, *The space complexity of pebble games on trees*, Inform. Process. Lett., 10 (1989), pp. 184–188.

[Me 78] F. MEYER AUF DER HEIDE, *A comparison between two variations of a pebble game on graphs*, Univ. of Bielefeld, Bielefeld, West Germany, 1978.

[PDS 77] G. PERSKY, D. DEUTSCH AND D. SCHWEIKERT, *LTX—a minicomputer-based system for automated LSI layout*, J. Design Automat. and Fault-Tolerant Comput., 1 (1977), pp. 217–255.

[Sh 79] Y. SHILOACH, *A minimum linear arrangement algorithm for undirected trees*, SIAM J. Comput., 8 (1979), pp. 15–32.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 1, March 1982

0196-5212/82/0301-0011 $01.00/0

SWITCHINGS CONSTRAINED TO 2-CONNECTIVITY IN SIMPLE GRAPHS*

R. TAYLOR†

Abstract. This paper follows as a natural extension of the ideas in a previous paper where we showed that any connected graph may be transformed by a sequence of switchings to any other connected graph of the same degree sequence, in such a way that all the intermediate graphs formed are connected. This was done for simple graphs, multigraphs and pseudographs. Here we show that the corresponding result is true for 2-connected simple graphs. The result for multigraphs and pseudographs will appear elsewhere. We also note that for k-connected graphs where $k \geqq 3$, this transformation theorem seems much more difficult to prove and in the last section of this paper we mention these difficulties.

1. Introduction. Unless otherwise specified, we will adopt the notation and terminology of Bondy and Murty [1].

The *degree sequence* of a graph denoted $\underline{d} = (d_1, \cdots, d_n)$ is the list of the degrees of all the vertices of the graph, conventionally arranged in nonincreasing order beginning with the maximum degree. When several of the terms in $\underline{d}$ are equal we may use exponential notation so that, for example, $\underline{d} = (3, 3, 2, 2) = (3^2, 2^2)$.

A graph G is a *realization* of a degree sequence $\underline{d}$ if the collection of the degrees of the vertices of G is the same as the collection of terms in $\underline{d}$. A *labelled* realization of a degree sequence $\underline{d} = (d_1, \cdots, d_n)$, $d_1 \geqq d_2 \geqq \cdots \geqq d_n$ is a graph whose vertices are labelled $v_1, \cdots, v_n$ with the restriction that the degree of v_i (denoted $d(v_i)$) is equal to d_i. Unless otherwise stated, any labelled realization of a degree sequence is to be considered as having this restriction.

A *switching* is a $\underline{d}$-invariant transformation on a graph that eliminates two edges and introduces two new ones. Thus a switching involves two edges (u, v) and (x, y) say, and transforms the graph by eliminating these edges and introducing new edges (u, x) and (v, y). This is illustrated in Fig. 1. Algebraically we may represent this operation as $[(u, v), (x, y)] \rightarrow [(u, x), (v, y)]$. Since we are dealing with simple graphs we may switch only when the edges (u, v) and (x, y) are independent, that is when u, v, x, and y are all different and the edges (u, x) and (v, y) are not already present in the graph.

To formalize the way in which the various realizations of a degree sequence are related by switchings, the following graph has been introduced by Eggleton and Holton [3].

DEFINITION. The *graph of realizations* (respectively, the graph of *labelled* realizations) of a degree sequence $\underline{d}$ is a graph whose vertices are identified with the

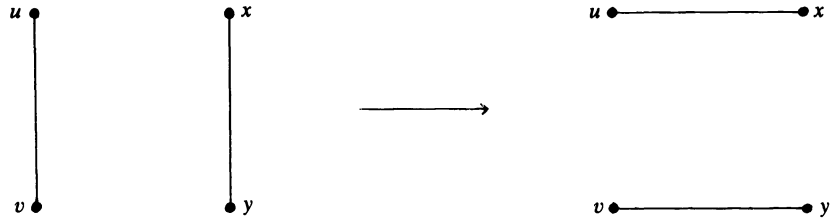


FIG. 1

* Received by the editors January 30, 1981, and in final form June 1, 1981.
† Department of Mathematics, University of Melbourne, Parkville, Victoria, 3052, Australia.

realizations (respectively, the labelled realizations) of $\underline{d}$, and where two vertices are adjacent if and only if the realizations corresponding to these vertices are a single switching apart. This is denoted by $R(\underline{d})$ and $R_l(\underline{d})$, respectively.

Often we are interested only in realizations of a degree sequence which have a certain property. Motivated by this we make the following definition.

DEFINITION. Let P be a property which a graph may possess. Then $R(\underline{d}, P)$, $R_l(d, P)$ are the subgraphs of $R(\underline{d})$, $R_l(\underline{d})$, respectively, induced by those vertices which correspond to graphs with property P.

Properties for which $R(\underline{d}, P)$ is connected are said to be *complete* (see Colbourn [2, p. 68]). If a property is complete we may find all the graphs of a given degree sequence with the property by switchings constrained to graphs with the property. Switching algorithms based on complete properties may provide a relatively efficient means of finding those realizations which possess the relevant property.

A relationship between completeness of a property in the labelled and unlabelled cases is given in the following theorem.

THEOREM 1.1. *Let P be a property of graphs. If $R_l(\underline{d}, P)$ is connected, then $R(\underline{d}, P)$ is connected.*

Proof. See [6]. □

Colbourn [2] showed that the property of being a tree is complete, and in [5] Syslo extended this to the property of being unicyclic. In [6] we generalized these results to show that the property of being connected is complete. The main result of this paper is that the property of being 2-connected is complete.

2. Safe switchings. We shall develop certain switching types which are used later in the proof of the main theorem.

DEFINITION. Let G be a 2-connected graph and σ a switching on G. We say the switching σ is *safe* if, when σ is applied to G, the resulting graph $\sigma(G)$ is also 2-connected.

DEFINITION. Let (x, y) and (u, v) be independent edges of a graph G. The switching $[(u, v), (x, y)] \to [(u, x), (v, y)]$ is called a *type*-1 switching if the edges (x, y) and (u, v) are on independent cycles in G.

LEMMA 2.1. *Any type-1 switching is safe.*

Proof. Let $[(u, v), (x, y)] \to [(u, x), (v, y)]$ be a type-1 switching on a 2-connected graph G that results in a graph G'. Assume that the switching is not safe and so G' is connected but not 2-connected (clearly the switching preserves connectivity). Thus G' must have some cut-vertex w. Let A be any component of $G' - w$ and B the rest of $G' - w$. Now since G was 2-connected, either (x, y) or (u, v) joins A and B (that is, has an end-vertex in A and in B). Suppose without loss of generality that $(x, y) \in E(G)$ is such an edge (see Fig. 2). Observe that w cannot be u or v since this would

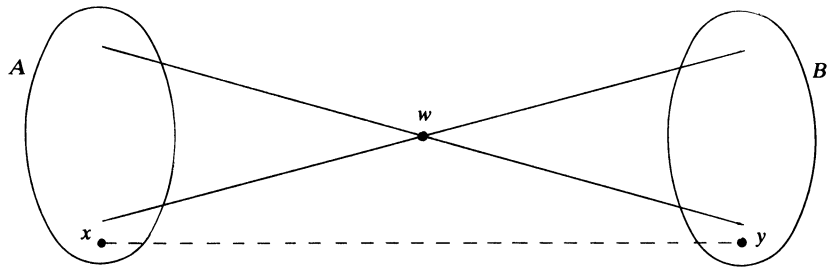


FIG. 2

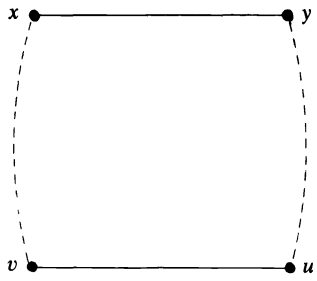


FIG. 3

imply that the edges (x, y) and (u, v) in G are not on independent cycles. For the same reason (u, v) cannot join A and B. Thus u and v must both lie in A or in B. Assume without loss of generality that u and v are in B. But then $(x, u) \in E(G')$ joins A and B, which contradicts the choice of w. □

DEFINITION. Let (x, y) and (u, v) be edges on a cycle C in G in such a way that there is a path between u and y on C independent of the vertices x and v, and a path between x and v independent of y and u (see Fig. 3). Under these circumstances we say the switching $[(u, v), (x, y)] \rightarrow [(u, x), (v, y)]$ is a *type-2* switching and that the cycle C is type-2 with respect to this switching.

LEMMA 2.2. *Any type-2 switching is safe.*

Proof. Let $\sigma : [(u, v), (x, y)] \rightarrow [(u, x), (v, y)]$ be a type-2 switching on a 2-connected graph G, and let $\sigma(G) = G'$. We assume that σ is not safe, and so G' is connected but not 2-connected. Thus G' must have a cut-vertex w and we may partition $G' - w$ into A and B as before. Further assume without loss of generality that (x, y) joins A and B, with, say, $x \in A$ and $y \in B$. Firstly we note that w cannot be u or v. To see this assume that $w = u$, say. Then since the edges (x, y) and (u, v) are on a cycle as in Fig. 3, we must have $v \in A$. But then (v, y) joins A and B, contradicting the choice of w. As in Lemma 2.1, we may also conclude that the vertices u and v cannot both be in A or in B. Now $(x, u) \in E(G')$ and $(y, v) \in E(G')$, and since there are no edges joining A and B in G', we must have $u \in A$ and $v \in B$ (see Fig. 4). By inspection we see that (u, v) and (x, y) cannot occur on any cycle in G in the required order. Thus σ is not a type-2 switching. This contradiction proves the lemma. □

3. 2-connectivity. In this section we prove that 2-connectivity is a complete property. This is done first for labelled graphs, the result for unlabelled graphs following as a corollary by Theorem 1.1.

THEOREM 3.1. *$R_l(\underline{d}, P)$ is connected, where $P \equiv$ "2-connected".*

Proof. The proof is by induction on n, the number of terms in $\underline{d}$.

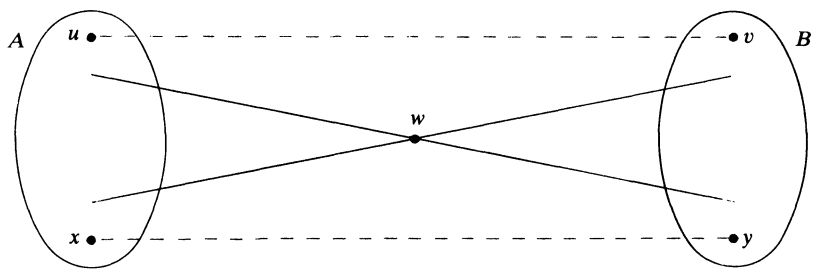


FIG. 4

$n = 3$. This follows trivially since there is only one labelled 2-connected simple graph on three vertices (see Fig. 5). Assume then that the result holds whenever $n \leqq m - 1$.

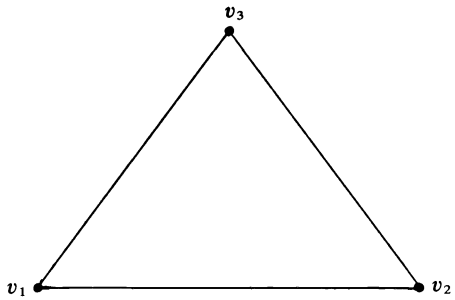


FIG. 5

$n = m \geqq 4$. First we treat the case $d = (2^m)$. The only 2-connected labelled realizations of (2^m) are isomorphic to C_m (the cycle on m vertices). Thus, we need only show that we may relabel a labelled C_m in any way by switchings which create only cycles on m vertices. To do this it is sufficient to indicate how any two adjacent vertices on the cycle may be interchanged. So let G be any labelled cycle on m vertices, and let $(\cdots, c, a, b, d, \cdots)$ be the labelled cycle G (see Fig. 6). We switch $[(c, a), (b, d)] \rightarrow [(c, b), (a, d)]$, and this interchanges the vertices a and b on the cycle.

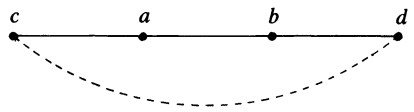


FIG. 6

Now for $\underline{d}$ to have any 2-connected realizations, clearly we must have $d_1 \geqq d_2 \geqq \cdots \geqq d_m \geqq 2$, and by the case just treated we may assume that $d_1 \geqq 3$. Further, we may also assume that $d_2 \geqq 3$ since the only connected realization of $\underline{d}$ with $d_1 \geqq 3$ and $d_2 = d_3 = \cdots = d_m = 2$ consists of a collection of cycles all passing through the vertex v_1 (see Fig. 7). Here v_1 is a cut-vertex, and so $\underline{d}$ has no 2-connected realizations. We proceed with the main body of the proof.

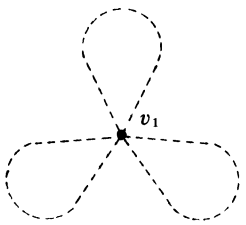


FIG. 7

Let G_1 and G_2 be any two labelled 2-connected realizations of $\underline{d}$. We show how to switch G_1 and G_2 into 2-connected graphs G_1^* and G_2^* in which $G_1^* - v_m$ and $G_2^* - v_m$ are both 2-connected graphs of the same degree sequence. This will allow us to use the induction result. Take G_1 for example. We shall switch on G_1 so as to

make v_m adjacent to $v_1, v_2, \cdots, v_s$ where $d(v_m) = d_m = s$. So let v_m be adjacent to $v_1, v_2, \cdots, v_p, v_{i_{p+1}}, \cdots, v_{i_s}$ in G_1, with $v_{p+1} \neq v_{i_j}$ for any $j > p$. We demonstrate how to switch on G_1 to make v_m adjacent to $v_1, \cdots, v_p, v_{p+1}, \cdots, v_{i_s}$. Now $d(v_{p+1}) \geqq d(v_{i_{p+1}})$ and so v_{p+1} must be adjacent to some vertex a which is not adjacent to $v_{i_{p+1}}$. Consider the edges $(v_m, v_{i_{p+1}})$, (a, v_{p+1}). Since G_1 is 2-connected they must both lie on some cycle. If the cycle is of the form indicated in Fig. 8, the type-2 switching $[(v_m, v_{i_{p+1}}),$

FIG. 8

$(v_{p+1}, a)] \rightarrow [(v_m, v_{p+1}), (v_{i_{p+1}}, a)]$ will give us the required result. Thus we may assume that the two edges lie on a cycle of the type shown in Fig. 9. Let b be the vertex adjacent to v_{p+1} on the path between v_{p+1} and v_m in Fig. 9. If b is not adjacent to $v_{i_{p+1}}$ the type-2 switching $[(v_m, v_{i_{p+1}}), (v_{p+1}, b)] \rightarrow [(v_m, v_{p+1}), (b, v_{i_{p+1}})]$ will produce a

FIG. 9

graph with the required properties. Assume then that b is adjacent to $v_{i_{p+1}}$ (see Fig. 10). But $d(v_{p+1}) \geqq d(v_{i_{p+1}})$ so v_{p+1} must be adjacent to some vertex $c \neq a$ which is not adjacent to $v_{i_{p+1}}$. Divide the subgraph shown in Fig. 10 into two parts, the cycle $(v_m, \cdots, b, v_{i_{p+1}}, v_m)$ denoted by C, and the rest of the graph denoted by R. Since G_1 is 2-connected there is a path between c and v_m independent of the vertex v_{p+1}. Let q be the first vertex on this path which is in the subgraph of Fig. 10. If $q \in R$, then

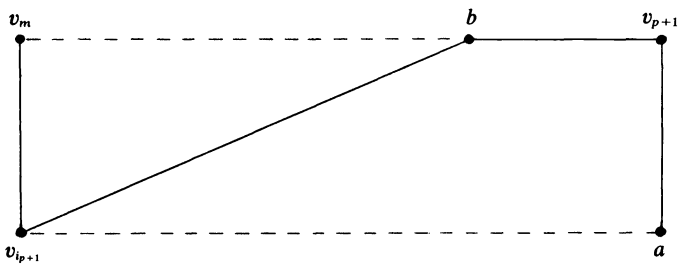


FIG. 10

$(v_m, v_{i_{p+1}})$ and (v_{p+1}, c) are on independent cycles, whilst if $q \in C$ and $q \neq v_{i_{p+1}}$, then (v_{p+1}, c) and $(v_{i_{p+1}}, v_m)$ are on a single cycle of type-2 with respect to the switching $[(v_{p+1}, c), (v_m, v_{i_{p+1}})] \rightarrow [(v_m, v_{p+1}), (c, v_{i_{p+1}})]$. This switching will be either of type-1 or type-2 and will result in a graph of the required form. Thus we may assume that $q = v_{i_{p+1}}$, and the situation is as depicted in Fig. 11. Now since $d(v_{p+1}) \geqq d(v_{i_{p+1}}) \geqq 4$ we must have yet another vertex d adjacent to v_{p+1} which is not adjacent to $v_{i_{p+1}}$. We partition the graph of Fig. 11 as we did in Fig. 10 into the cycle C and the remainder R' and by the same reasoning as above we may assume that there is a path from d to $v_{i_{p+1}}$ independent of any other vertex in Fig. 11. But again $d(v_{p+1}) \geqq d(v_{i_{p+1}}) \geqq$

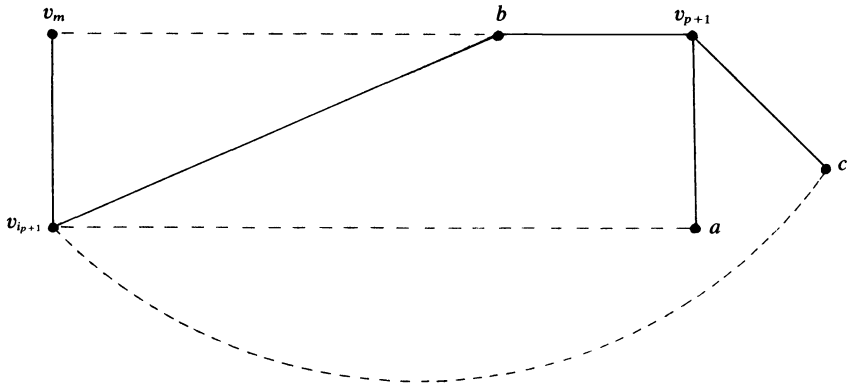


FIG. 11

5 and we continue in this way until we find some vertex e adjacent to v_{p+1}, not adjacent to $v_{i_{p+1}}$, where (v_{p+1}, e) and $(v_{i_{p+1}}, v_m)$ are both on independent cycles or on a single cycle of type-2 with respect to the switching $[(v_m, v_{i_{p+1}}), (v_{p+1}, e)] \rightarrow [(v_m, v_{p+1}), (v_{i_{p+1}}, e)]$. In either case this switching is safe and produces a graph in which v_m is adjacent to $v_1, \cdots, v_p, v_{p+1}, v_{i_{p+2}}, \cdots, v_{i_s}$ as desired. Continuing this process we transform G_1 to G_1' by a sequence of safe switchings where v_m is adjacent to $v_1, v_2, \cdots, v_s$ in G_1'.

We now perform a series of safe switchings on G_1' and transform it into a graph G_1^* where $G_1^* - v_m$ is 2-connected and the neighborhood of v_m is unaffected. So assume $H = G_1' - v_m$ is not 2-connected and has cut-vertices $x_1, \cdots, x_k$ which separate connected blocks $A_1, \cdots, A_t$, where each A_i contains no cut-vertices of H. We observe that the graph H has at least two blocks which are adjacent to at most one cut-vertex each. These blocks together with their respective cut-vertices correspond to end-vertices in the block-cutpoint tree of H (see Harary [4, p. 36]). We now describe how to switch on H so as to decrease the number of blocks by at least one.

Let A_i and A_j be any two blocks of H which are adjacent to only one cut-vertex each. Note that every vertex adjacent to v_m in G_1' has degree at least three. This follows for if $d_m = 2$, then v_n is adjacent only to v_1 and v_2 which both have degree at least three, whilst if $d_m \geqq 3$, then every vertex has degree at least three. Now since G_1' is 2-connected v_m must be adjacent to a vertex in A_i and a vertex in A_j, say x and y, respectively. Since $d(x) \geqq 3$ and $d(y) \geqq 3$, x and y must have degree at least one in A_i and in A_j. And so A_i, A_j must contain edges $(x, z) \in E(A_i)$ and $(y, w) \in E(A_j)$. By construction x, z, y, w are not cut-vertices of H and so (x, z) and (y, w) are on cycles in H. If (x, z), (y, w) are on independent cycles in H, then we make the type-1 switching $[(x, z), (y, w)] \rightarrow [(x, y), (z, w)]$ which merges A_i and A_j into one larger block

and so decreases the number of blocks in H by at least one. Assume on the other hand that (x, z) and (y, w) are not on independent cycles in H. This could only occur if they are separated by a single cut-vertex x_r with both cycles passing through x_r. Thus x_r is adjacent to at least two vertices in A_i and in A_j. Now if there is a cycle in A_i or A_j, then any edge on that cycle and any edge in the other block must be on independent cycles. Switching between these edges is safe of type-1 and results in a graph which has at least one block fewer than H has. We may therefore assume that both A_i and A_j are trees. In the following we consider various cases for the value of d_m, and show that each leads to a contradiction.

$d_m = 2$. Here we must have $x = v_1$ and $y = v_2$ (see Fig. 12). We now show that x_r is adjacent to at least $d_1 - 1$ vertices of A_i. If x_r is not adjacent to v_1, then the degree

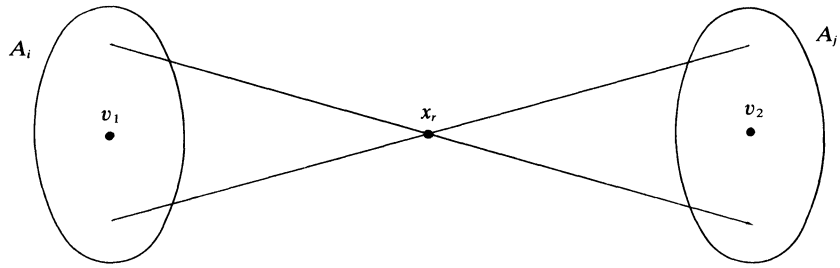


FIG. 12

of v_1 in A_i is $d_1 - 1$ and so A_i has at least $d_1 - 1$ vertices of degree 1. Now in H all vertices of A_i are not cut-vertices and so every vertex of degree 1 in A_i must be adjacent to x_r in H. Thus x_r is adjacent to at least $d_1 - 1$ vertices in A_i. If on the other hand x_r is adjacent to v_1, then the degree of v_1 in A_i is $d_1 - 2$, and so using the argument above we may conclude that x_r is adjacent to at least $d_1 - 2$ vertices of degree 1 in A_i (other than v_1). Including v_1 we see that x_r is adjacent to at least $d_1 - 1$ vertices of A_i. Thus $d(x_r) \geqq d_1 - 1 + 2 = d_1 + 1$, and this is impossible.

$d_m = 3$. In this case v_m must have only one edge in common with either A_i or A_j, say A_i. Since A_i is a tree it has at least two vertices of degree one, let q be one of them which is not adjacent to v_m. But in G_1' we can have at most one extra edge incident with q (between it and x_r) and so $d(q) \leqq 2$. But this contradicts the fact that $d(q) \geqq d_m = 3$.

$d_m \geqq 4$. Clearly all vertices of A_i and A_j have degree at least two and so cannot be trees.

$$G_1 \xrightarrow{\sigma_1} G_1^* \xrightarrow{\theta} G_2^* \xleftarrow{\sigma_2} G_2$$

FIG. 13

Thus we may switch G_1' to a 2-connected graph G_1'' where the neighborhood of v_m is unaltered and where $G_1'' - v_m$ has at least one block fewer than $G_1' - v_m$. It follows then that by a sequence of safe switchings we may transform G_1' into a graph G_1^* where v_m is adjacent to $v_1, v_2, \cdots, v_s$ in G_1^* and $G_1^* - v_m$ contains only one block and so is 2-connected. Thus we may transform G_1 into G_1^* by a sequence of safe switchings σ_1. Similarly by a sequence of safe switchings σ_2 we may transform G_2 into a graph G_2^* with v_m adjacent to $v_1, v_2, \cdots, v_s$ in G_2^*. Now $G_1^* - v_m$ and $G_2^* - v_m$ are both 2-connected graphs with degree sequences $\underline{d}' = (d_1 - 1, \cdots, d_s - 1, d_{s+1}, \cdots, d_{m-1})$, and so by the induction hypothesis we may trans-

form $G_1^* - v_m$ into $G_2^* - v_m$ by a sequence of safe switchings θ. Thus the sequence of switchings defined by $\sigma_1\theta\sigma_2^{-1}$ (acting from the left) transforms G_1 into G_2 in the required manner (see Fig. 13). Since G_1 and G_2 were arbitrary 2-connected labelled realizations of $\underline{d}$ we have the result for $n = m$. The theorem follows by induction. □

COROLLARY 3.1. *$R(\underline{d}, P)$ is connected, where $P \equiv$ "2-connected".*

Proof. By Theorem 1.1. □

4. k-connectivity. In view of the fact that k-connectivity is a complete property for $k = 1, 2$, we believe there are sufficient grounds to suppose that the same result is true for all k. In this section we indicate some of the difficulties involved in trying to prove this using the same basic approach that has been successful for $k = 1, 2$. Take $k = 3$ for example.

(i) We must be able to switch on a 3-connected graph so as to make v_n adjacent to $v_1, \cdots, v_s$, where $d_n = s$. To ensure that the graphs formed at each step are 3-connected we would have to develop switching types that preserve 3-connectivity.

(ii) We need to switch on a 3-connected realization G of a degree sequence $\underline{d}$ so that $G - v_n$ is 3-connected. However this is clearly not possible for any $\underline{d}$ with $d_3 \leqq 3$ since then $G - v_n$ would have at least one vertex of degree 2. This class contains for instance all the 3-connected cubic graphs. Thus the case $\underline{d} = (3^n)$ would have to be treated as a special case just as we treated $\underline{d} = (2^n)$ for 2-connectivity (see Theorem 3.1). However, although the structure of the 2-connected two regular graphs is elementary (they are simply cycles), there are many nonisomorphic 3-connected cubic graphs on n vertices.

As a first step then in proving that 3-connectivity is a complete property, we believe a proof showing that this property is complete for cubic graphs would be useful. The techniques used in such a proof may even lead to a solution of the general problem concerning k-connectivity.

REFERENCES

[1] J. A. BONDY AND U. S. R. MURTY, *Graph Theory with Applications*, Macmillan, London, 1976.

[2] C. J. COLBOURN, *Graph generation*, Res. Rep. CS-77-37, Computer Science Dept., University of Waterloo, Waterloo, Ontario, Canada, November 1977.

[3] R. B. EGGLETON AND D. A. HOLTON, *Graphic sequences*, Combinational Mathematics VI, Proc. 6th Australian Conference, Lecture Notes in Mathematics 748, Springer-Verlag, New York, 1979, pp. 1–10.

[4] F. HARARY, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.

[5] M. M. SYSLO, *On tree and unicyclic realizations of degree sequences*, Research Report CS-80-064, Computer Science Dept., Washington State University, Pullman, July 1980.

[6] R. TAYLOR, *Constrained switchings in graphs*, to appear.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 1, March 1982

0196-5212/82/0301-0012 $01.00/0

A CLASS OF MATRICES CONNECTED WITH VOLTERRA PREY-PREDATOR EQUATIONS*

RAY REDHEFFER† AND ZHOU ZHIMING‡

Abstract. If (p_{ij}) is a real $n \times n$ matrix, conditions are given which ensure $(a_i p_{ij}) \leqq 0$ for some set of positive constants a_i, where $(a_i p_{ij}) \leqq 0$ means that the associated quadratic form is nonpositive. Within this class, a companion paper gives an effectively complete solution to the problem of asymptotic stability for n-variable Volterra prey-predator equations; that is why the class is important. Both the algebraic problem considered here and the analytic problem of stability involve a novel blend of topological and combinatorial considerations.

1. Introduction. Throughout this paper $p = (p_{ij})$ is a real $n \times n$ matrix such that $p_{ii} \leqq 0$ for $i = 1, 2, \cdots, n$ and

$$p_{ij}p_{ji} < 0 \quad \text{if } (i-j)p_{ij} \neq 0, \quad i, j = 1, 2, \cdots, n.$$

A *perturbation* of p is a change to another matrix $\tilde{p}$ such that $\tilde{p}_{ij} = 0$ if, and only if, $p_{ij} = 0$. The perturbation is small if $\max|\tilde{p}_{ij} - p_{ij}|$ is small. Our objective is to give conditions under which p is *stably admissible* in the sense of the following definition:

DEFINITION 1. The matrix p is *admissible* if there exist positive constants a_i such that $(a_i p_{ij}) \leqq 0$, where this condition means

$$\sum_{i,j=1}^{n} a_i p_{ij} w_i w_j \leqq 0 \quad \text{for all } (w_1, w_2, \cdots, w_n) \in \mathbb{R}^n.$$

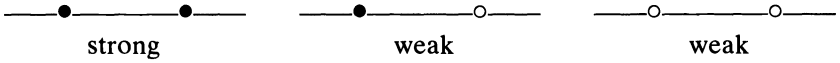
The matrix is *stably admissible* if every sufficiently small perturbation $\tilde{p}$ is admissible.

These conditions have an interesting bearing on the Volterra system

$$\dot{x}_i = x_i\left(e_i + \sum_{j=1}^{n} p_{ij}x_j\right), \qquad i = 1, 2, \cdots, n. \tag{1}$$

It turns out that the study of boundedness and global asymptotic stability of solutions is very much simplified if p is admissible. The condition of stable admissibility is more appropriate than admissibility, however, because the coefficients are known only with limited precision. The relation of Definition 1 to (1) is fully discussed in [6] and motivates the present investigation. See also Volterra [7, Ch. III], where the importance of the two cases (i) p skew symmetric and (ii) $p_{ii} < 0$ for all i is convincingly demonstrated.

As in [6], the *graph* of p is a graph of n vertices or nodes $1, 2, \cdots, n$ in which the node i is adjacent to j in the sense of [1] if and only if $p_{ij} \neq 0$. The graph has a black dot at i if $p_{ii} < 0$ and an open circle ○ if it is known only that $p_{ii} \leqq 0$. We assume that the graph is connected, since if it is not, the problem breaks up into two or more simpler problems of the same type. An edge which directly connects two black dots is called a *strong link*, and otherwise the link is weak, thus:



strong weak weak

* Received by the editors September 15, 1980, and in revised form June 1, 1981.

† Department of Mathematics, University of California, Los Angeles, California 90024. The work of this author was supported in part by the National Science Foundation under grant MCS 79-03544.

‡ Department of Mathematics, University of California, Los Angeles, California 90024. This author is an exchange scholar under the auspices of the exchange program of the University of California, Los Angeles, and Zhongshan University, Canton, China.

We use $(1, 2, \cdots, m)$ or $[1, 2, \cdots, m]$ to denote a path or loop, respectively, which connects the nodes $1, 2, \cdots, m$ in succession. (The latter has an edge joining 1 to m, the former does not). Aside from these special conventions borrowed from [6], we use standard graph-theoretic terminology as described, for example, in [1]. In this terminology the graph of p is *connected*, *unoriented*, and except for the black dots, *unlabeled*.

If $\tilde{p}$ is a sufficiently small perturbation, p and $\tilde{p}$ have the same graph, including the distribution of black dots. This is the reason for restricting the term *perturbation* so that no new nonzero coefficients are introduced, and no nonzero coefficients are changed to 0. When the perturbation is sufficiently small, the inequalities $p_{ij}p_{ji} < 0$ for $i \neq j$ are also preserved.

We conclude this introductory discussion with a brief summary of the historical development. The most important reference is Volterra [7]. Here we find the basic hypothesis of sign antisymmetry, the introduction of the multipliers a_i, and the fundamental condition of equality for products of successive coefficients (see eq. (7) below). Sign antisymmetric matrices are discussed further in [3] and [5], and as was pointed out by the referee, the condition for diagonalization there developed gives an alternative approach to the results of Volterra. Additional topics in the same circle of ideas are taken up in [4], where matrices satisfying Volterra's hypothesis $p_{ij} \neq 0 \Rightarrow p_{ji} \neq 0$ are referred to as *combinatorially symmetric*. The fact that none of these references mentions [7] perhaps lends color to Krikorian's remark [2] that Volterra's work is still insufficiently known and improperly understood.

Volterra did not introduce the graph of p and makes no distinction between trees and loops. (In the case of a path free of loops his condition holds automatically, since both products are 0). However, aside from our labeling, the graph is introduced, in just the form used here, in [5]; in [2], [3], by contrast, the graph is *directed* after the manner of current research in control theory.

2. General remarks. If the graph of p is a tree, then p is admissible since a_i can be chosen so that

$$a_i p_{ij} + a_j p_{ji} = 0 \tag{2}$$

whenever i and j are adjacent. Conversely, this condition is necessary for admissibility if i and j are adjacent and (i, j) is not a strong link. Hence, every loop must contain at least one strong link if p is stably admissible. This fact plays a basic role in [6].

In applications it is essential to choose $a_i > 0$ not only so that $(a_i p_{ij}) \leqq 0$ but so that, in addition,

$$\sum_{i,j=1}^{n} a_i p_{ij} w_i w_j = 0 \Rightarrow p_{ii} w_i = 0, \qquad i = 1, 2, \cdots, n.$$

Such a choice of a_i may not be possible if p is merely admissible but is always possible when p is stably admissible To see this, let ε be a small positive constant and let

$$\tilde{p}_{ij} = p_{ij} \quad \text{for } i \neq j, \qquad \tilde{p}_{ii} = (1 - \varepsilon) p_{ii}.$$

If a_i are so chosen that $a_i > 0$ and $(a_i \tilde{p}_{ij}) \leqq 0$ the equation

$$\sum a_i p_{ij} w_i w_j = \sum a_i \tilde{p}_{ij} w_i w_j + \varepsilon \sum a_i p_{ii} w_i^2$$

shows that the quadratic form on the left is $\leqq 0$ and that it can vanish only if $p_{ii} w_i = 0$ for each i.

From another point of view it will be found that $(a_i p_{ij}) \leqq 0$ leads to a certain number of weak inequalities, that is, inequalities of the form $A \leqq B$. If we require these same inequalities to be strong, $A < B$, we can split off εp_{ii} as above and again get the side condition $p_{ii} w_i = 0$. Note also that strong inequalities are preserved by small perturbations.

Another remark of a general nature pertains to the uniting of graphs. Suppose two matrices p and p^* are stably admissible, and suppose the graph for $p + p^*$ is obtained by joining the graphs for p and for p^* at a single point; in other words, the graphs for p and p^* have exactly one vertex in common. Then $p + p^*$ is stably admissible. This is so because the conditions involving the a_j are homogeneous. Let a_j be the coefficients for p and a_j^* for p^*. Then λa_j^* will also do for p^*, where λ is any positive constant; and we can choose λ so that $\lambda a_i^* = a_i$ at the particular vertex i which the graphs have in common.

The same method gives the following: Suppose the graphs for p and p^* have no vertex in common, and form a new graph by connecting one vertex i of p to one vertex j of p^*. Then the matrix corresponding to the new graph is stably admissible if, and only if, p and p^* are stably admissible; we assume that the new coefficients introduced by this process satisfy $p_{ij} p_{ji} < 0$. For proof, choose $\lambda > 0$ so that

$$a_i p_{ij} + \lambda a_j^* p_{ji} = 0$$

and reason as above. Necessity follows because the variables associated with p or p^* could be taken to be 0, while the others remain arbitrary.

As a special case, these remarks show that adding any number of trees to the graph of a stably admissible matrix (without forming any new loops) will again lead to a stably admissible matrix. Thus, in getting conditions for admissibility, trees can be ignored.

3. The case $n = 3$. In [7] it is seen that a necessary and sufficient condition for admissibility when $n = 3$ can be deduced, theoretically, from two homogeneous inequalities of degrees 13 and 26, respectively. Here we give a simpler criterion, based on the Hurwitz inequalities, which reduces the problem effectively to the solution of a single quartic equation. The discussion sheds light on the difficulties associated with the general case.

Since the problem is trivial for a tree, let the graph with $n = 3$ be a triangle. If there is no strong link the matrix cannot be stably admissible, and hence we assume $p_{11} p_{22} \neq 0$. A sufficient condition for stable admissibility is then

$$R + \frac{1}{R} - 2 < 4 \frac{p_{11} p_{22}}{|p_{12} p_{21}|}, \quad \text{where } R = \frac{|p_{12} p_{23} p_{31}|}{|p_{21} p_{32} p_{13}|}. \tag{3}$$

For the proof, it is readily checked that the condition is both necessary and sufficient when $p_{33} = 0$, and a term in p_{33} is helpful.

If all three diagonal entries p_{ii} are negative one can state three conditions like the above, and any one of them is sufficient. To get a condition which is both necessary and sufficient (when each $p_{ii} < 0$) assume without loss of generality that $p_{ii} = -1$ and $a = (1, x, y)$. Denoting the corresponding matrix $(a_i p_{ij})$ by $A(x, y)$ we see that p is stably admissible if and only if for some positive (x, y)

$$M(x, y) = A(x, y) + A^t(x, y) < 0 \tag{4}$$

where M is defined by the equation. We set $D(x, y) = \det M(x, y)$. Then a necessary and sufficient condition for (4) is $D(x, y) < 0$ and also

$$4x > (p_{12} + x p_{21})^2, \quad 4y > (p_{13} + y p_{31})^2, \quad 4xy > (x p_{23} + y p_{32})^2. \tag{5}$$

The inequalities (5) reduce to

$$x_0 < x < x_1, \quad y_0 < y < y_1, \quad m_0 < \frac{y}{x} < m_1, \tag{6}$$

where x_i, y_i and m_i are roots of the quadratics obtained with = instead of $<$ in (5). The hypothesis $p_{ij}p_{ji} < 0$ ensures that these roots are *positive* and *unequal*. Their values are regarded as known.

The region defined by (6) is called the *feasible region* and is illustrated in three typical cases in Fig. 1. A necessary and sufficient condition for stable admissibility is that

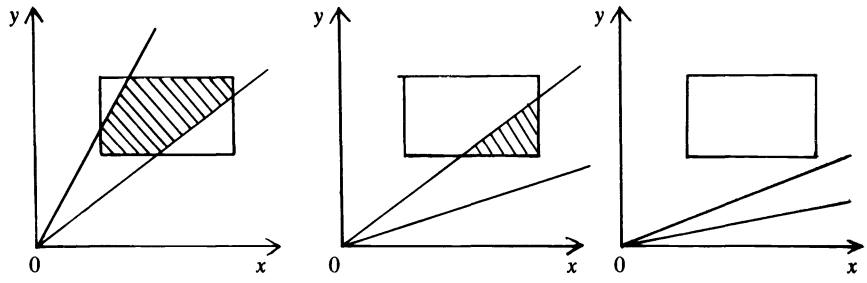


FIG. 1

the feasible region be nonempty and that $D(x, y) < 0$ at some point thereof. By a brief analysis it is found that evaluation of $D(x, y)$ at 12 points (x_i, y_i) suffices to decide the question. Of these points eight (on the boundary of the rectangle) are obtained from quadratic equations while four (interior) require a quartic. Evaluation of $D(x, y)$ at 12 specified points lends itself to automatic computation, or alternatively, and perhaps preferably, one could apply any standard minimizing algorithm to $D(x, y)$ over the feasible region.

4. A general criterion. According to [6] the loop $[12 \cdots m]$ is *balanced* if it satisfies Volterra's condition,

$$|p_{12}p_{23} \cdots p_{m1}| = |p_{21}p_{32} \cdots p_{1m}|. \tag{7}$$

(Similar products were introduced in another connection in [4].) We want a *measure of asymmetry* to describe the extent to which this condition fails. If (p_{ij}) is replaced by $(c_{ij}p_{ij})$ where c is a symmetric matrix with nonzero elements, it is desired that the measure of asymmetry shall remain unchanged. This suggests the ratio R of the two quantities in (7) as a measure of asymmetry. However, we also want the measure of asymmetry to be independent of the direction in which the path is traversed. Since reversing the direction changes R into $1/R$, the measure $R + 1/R$ is suggested. Finally, we would like the measure to be 0 when the loop is balanced, that is, when (7) holds.

These remarks may serve to motivate the following definition:

DEFINITION 2. The measure of asymmetry of the loop $[12 \cdots m]$ is

$$A = R + \frac{1}{R} - 2 \quad \text{where } R = \frac{|p_{12}p_{23} \cdots p_{m1}|}{|p_{21}p_{32} \cdots p_{1m}|}.$$

A similar definition applies to any loop, with a more elaborate use of subscripts. It should be noticed that the value of A is not affected if we retrace steps in traversing the loop, going backward and then again forward. The extra coefficients p_{ij} introduced by

the backtracking occur in both numerator and denominator of R and cancel out. Further properties of this sort are discussed in § 6.

We also require the following:

DEFINITION 3. The strength of a strong link joining i directly to j is measured by

$$B_{ij} = \frac{p_{ii}p_{jj}}{|p_{ij}p_{ji}|}.$$

Suppose next that a graph is such that every loop in it has at least one strong link. Assuming that at least one loop is actually present, choose a loop and break one of its strong links. If a loop is still present, choose another loop and break one of its strong links, and so on. Continuing this process, we get a definite set of strong links such that breaking all of them reduces the graph to a tree.

Underlying this procedure is a theorem which in the referee's formulation reads as follows: Suppose G is a connected graph with the property that every cycle has at least one distinguished line. Then the successive removal of these lines by the above rules always results in a spanning tree T of G. Moreover the lines so removed constitute a set of chords of the tree T and the set of cycles from which they were removed is a basis for the cycles of G. A formal proof follows by induction on the number E of edges. For $E = 3$ the result is obvious, and removing a single one of the distinguished lines reduces the case $E+1$ to the case E.

We specify a strong link by giving its endpoints; thus, L_{ij} is a strong link joining i to j. The link is not oriented, so that the same link is represented by L_{ji}. Let us list only the links that were used in the above process, and each link only once. That being done, let $n(i)$ be the number of times the index i occurs in this list as a subscript on L. For example, if the links in the list are

$$L_{12}, \quad L_{34}, \quad L_{13}, \quad L_{41}, \quad L_{23}, \quad L_{35},$$

then $n(1) = 3$, $n(2) = 2$, $n(3) = 4$, $n(4) = 2$, $n(5) = 1$.

Since the graph after removal of the strong links L_{ij} is a tree, there is a unique simple path from i to j after the removal. Restoring the single link L_{ij} we get a unique measure of asymmetry A_{ij} associated with the nodes i and j. A measure of the strength of L_{ij} is given by B_{ij} in Definition 3.

The following theorem is the principal goal of this discussion:

THEOREM 1. *Let L_{ij} be strong links whose removal generates a tree as described above. Then the matrix p associated with the original graph is stably admissible if the inequality*

$$A_{ij} < 4\frac{B_{ij}}{n(i)n(j)}$$

holds for each pair of subscripts (i, j) in the set $\{L_{ij}\}$.

5. Proof of Theorem 1. For a single loop $[12 \cdots m]$ with a strong link at $(1, m)$ Theorem 1 is established by choosing $a_i > 0$ so that

$$a_i p_{i\,i+1} + a_{i+1}p_{i+1\,i} = 0, \qquad i = 1, 2, \cdots, m-1. \tag{8}$$

Then all cross products $w_i w_{i+1}$ except $w_m w_1$ disappear from the associated quadratic form and the latter reduces to

$$\sum_{i=2}^{m-1} a_i p_{ii} w_i^2 + a_1 p_{11} w_1^2 + (a_1 p_{1m} + a_m p_{m1}) w_1 w_m + a_m p_{mm} w_m^2.$$

Clearly a sufficient condition for stable admissibility is

$$(a_1 p_{1m} + a_m p_{m1})^2 < 4 a_1 a_m p_{11} p_{mm} \tag{9}$$

and this condition is necessary as well as sufficient if $(1, m)$ is the only strong link in the loop. Writing (8) in the form

$$\frac{a_{i+1}}{a_i} = \left| \frac{p_{i\,i+1}}{p_{i+1\,i}} \right|$$

we find, by an elementary computation, that (9) is equivalent to $A_{1m} < 4B_{1m}$ in the notation of Theorem 1.

It is important that any particular value a_i in this process can be arbitrarily prescribed, subject to $a_i > 0$, and then the remaining a_j are uniquely determined. Because of that, the process applies to several loops simultaneously, provided their strong links are disjoint. The latter condition means $n(i) = 1$ in the notation of Theorem 1.

Suppose, then, that the removal of a certain set of disjoint strong links L_{ij} leads to a tree. Let a_i be determined for the tree by the equations analogous to (8), so that all cross products except those associated with the L_{ij} vanish. (The details require a more elaborate notation which, however, will not be spelled out here.) The important feature is that, if we look at the loop associated with any given strong link, the definition of a_i for the tree agrees with the definition of a_i for the loop that led to the above condition $A_{1m} < 4B_{1m}$. Hence, if $A_{ij} < 4B_{ij}$ holds for each link L_{ij}, we find again that p is stably admissible. The condition is necessary and sufficient if the only strong links in the graph are the L_{ij}.

Finally, we have to account for the possibility that $n(i) > 1$. This means that the index i is involved in more than one strong link, and hence, the term $a_i p_{ii} w_i^2$ must be suitably parceled out. To this end let us write

$$p_{ii} = \frac{p_{ii}}{n(i)} + \frac{p_{ii}}{n(i)} + \cdots + \frac{p_{ii}}{n(i)} \qquad (n(i) \text{ terms})$$

and use one of these terms for each of the strong links involving the index i. This has the effect of replacing p_{ii} by $p_{ii}/n(i)$ in the foregoing calculation, and likewise, p_{jj} is replaced by $p_{jj}/n(j)$. The earlier condition $A_{ij} < 4B_{ij}$ then becomes the same as the condition in Theorem 1. □

It is clear that the process used for implementation of Theorem 1 is far from unique, and that one sequence of removed links may fail to show stable admissibility while another sequence succeeds. This possibility was already illustrated in the case $n = 3$, where we pointed out that one can permute subscripts in (3) if each $p_{ii} < 0$.

6. Balanced matrices. If the graph contains a loop, the requirement that a matrix be balanced is not preserved by small perturbations, as we have already remarked. Nevertheless, balanced matrices have many interesting properties and a brief discussion is given now. The graph of a balanced matrix p will be termed *p-balanced.* In this context it is thought that the graph is oriented and that it has the label $|p_{ij}/p_{ji}|$ attached to the edge from i to j.

For an open path $(1, 2, \cdots, m)$ we define a ratio

$$R^* = \frac{|p_{12} p_{23} \cdots p_{m-1\,m}|}{|p_{21} p_{32} \cdots p_{m\,m-1}|}$$

analogous to R; in fact $R^* = R|p_{1m}/p_{m1}|$. One of the most interesting properties of

p-balanced graphs can now be formulated as follows. Let i and j be any two distinct vertices of a p-balanced graph and let L be any path whatever in the graph which starts at i and ends at j. (It is permitted that L have arbitrarily many loops and self-intersections.) Then the ratio R^* computed relatively to L is independent of L; it depends on the ordered pair (i, j) only. The proof is left to the reader.

This remarkable fact leads to a far-reaching generalization of Theorem 1. Instead of a tree, let us start with any p-balanced graph, and introduce strong links L_{ij} at selected nodes (i, j) (see dotted lines in Fig. 2). By the above remark R_{ij}^* is well defined,

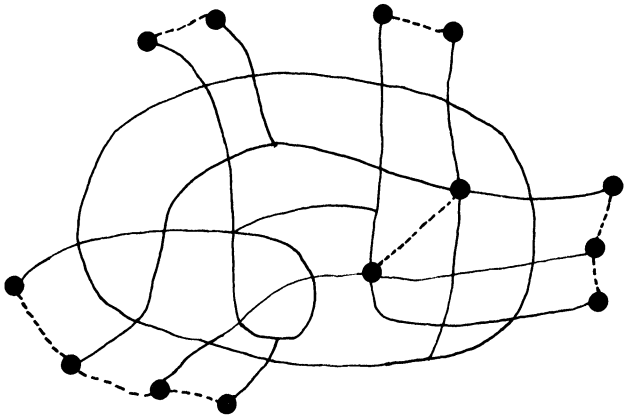


FIG. 2

hence R_{ij} is well defined, and therefore A_{ij} is. When these new links L_{ij} are introduced the new network is not, as a rule, balanced. But it will be stably admissible if each loop in the original graph has a strong link and if

$$A_{ij} < 4\,\frac{B_{ij}}{n(i)n(j)}$$

for each of the added links L_{ij}. This follows as in the proof of Theorem 1.

In the special case of Theorem 1, the p-balanced graph with which the process starts is a tree and A_{ij} is well determined because there is a unique path from i to j that does not intersect itself (actually backtracking does no harm, as stated in § 4). Here, on the contrary, there may be a multiplicity of paths, and the requirement that the matrix be balanced is essential. In fact, if R_{ij}^* is independent of path for a *single pair* (i, j), it follows necessarily that the graph is p-balanced.

7. Analogy to electrical networks. The fact that R^* is independent of the path is reminiscent of the theory of electrical networks, where the voltage drop from node i to node j is also independent of the path. Upon a change in orientation of the path R^* is changed to its reciprocal and the voltage to its negative. For quantitative development of this analogy, let the voltage drop from i to j be defined by

$$V_{ij} = \log \frac{|p_{ij}|}{|p_{ji}|}.$$

Then a matrix is p-balanced if and only if the total voltage drop around any closed loop in its graph is 0. When this holds we can assign voltages V_i at the ith node and determine

V_{ij} by

$$V_{ij} = V_i - V_j.$$

The choice of any particular voltage, say V_1, as reference level is arbitrary. The condition that the graph be p-balanced is precisely the condition that this representation for V_{ij} shall lead to no inconsistency.

We shall not pursue this analogy in detail here, but mention that it gives

$$n + \frac{n(n-1)}{2} + (n-1) = \frac{n^2+3n}{2} - 1$$

for the total number of independent parameters in an $n \times n$ balanced matrix. The same result is obtained by introduction of triangles as a basis for the loops of the graph, as the reader can verify.

8. The clam shell. There are two respects in which Theorem 1 can be improved. First, one can use a more general decomposition of p_{ii} such as

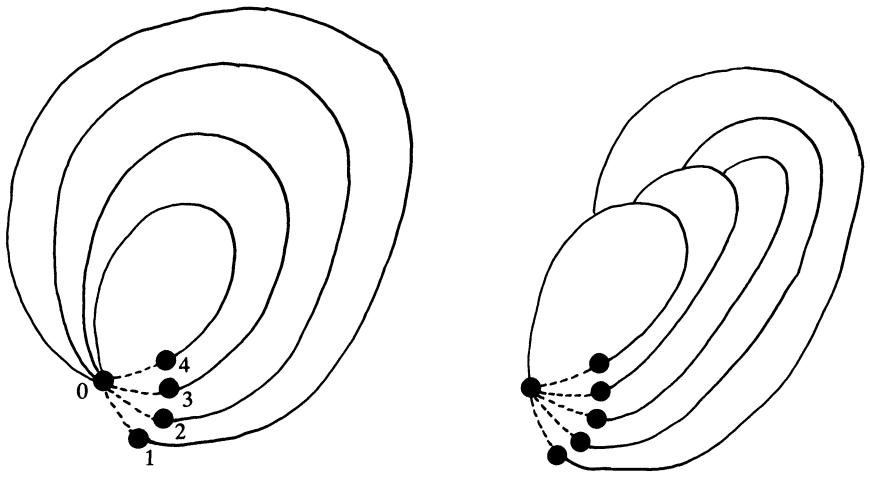
$$p_{ii} = \sum_{k=1}^{n(i)} c_{ik} p_{ii}, \qquad c_{ik} > 0, \quad \sum_{k=1}^{n(i)} c_{ik} = 1.$$

Theorem 1 corresponds to the choice $c_{ik} = 1/n(i)$. Second, one can exploit any strong links other than the L_{ij} that may be present. We shall illustrate these refinements in several examples. The scope of these examples is increased by the possibility of combining graphs, as explained in § 2.

As a first example, let the links L_{ij} be such that they have one end in common and the other ends are all distinct. The common end is labeled $i = 0$ and the other ends $i = 1, 2, \cdots, m$ as shown for $m = 4$ in Fig. 3a. This figure resembling a clam shell is the simplest example of the type of graph we have in mind and suggests the name. However, the configuration can be generalized as in Fig. 3b. The main requirement is that breaking all of these strong links produces a tree. We write

$$P_{00} = (c_1 + c_2 + \cdots + c_m) p_{00}$$

where the c_i are positive constants whose sum is 1.



(a) (b)

FIG. 3

If A_i is the measure of asymmetry associated with the ends $(0, i)$ of the tree, and B_i is the strength of the strong link joining 0 to i in the original graph, the condition $A_i < 4c_iB_i$ ensures that the matrix p is stably admissible. The constants c_i can be determined if, and only if,

$$\frac{A_1}{B_1}+\frac{A_2}{B_2}+\cdots+\frac{A_m}{B_m}<4.$$

This condition represents a very substantial improvement on Theorem 1, since the latter requires $A_i/B_i < 4/m$ for each i.

9. The ladder. We suppose now that $n(i) = 2$ except for $i = 0$ and $i = m$, in which case $n(i) = 1$. The typical graph satisfying these conditions is obtained by closing links in a comb, as suggested by the dotted lines in Fig. 4. The graph obtained after closure looks like a ladder, hence the name.

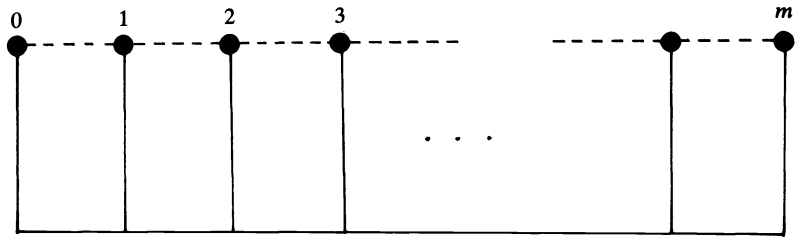


FIG. 4

As indicated by the figure, the free ends of the tree are numbered consecutively from 0 to m. We write

$$p_{ii} = (1-c_i)p_{ii} + c_ip_{ii}, \qquad 0<c_i<1,$$

for $i = 1, 2, \cdots, m-1$. If A_i is the measure of asymmetry for the small loop containing $(i-1, i)$ and B_i is the strength of the link joining $(i-1, i)$ a sufficient condition for stable admissibility is

$$M_1<1-c_1, \quad M_2<c_1(1-c_2), \quad M_3<c_2(1-c_3), \quad \cdots, \quad M_m<c_{m-1},$$

where $M_i = A_i/(4B_i)$. The conditions on c_i are least restrictive when c_{i-1} is as large as possible. Hence, the optimum choice is obtained if we choose c_i to give equality in all relations except the last and strict inequality in the last. By a small change of $\{c_i\}$ the leeway given in the last relation can be distributed over the others, so that strict inequality holds in all of them.

This procedure show that p is stably admissible if the m quantities

$$M_1, \quad \frac{M_2}{1-M_1}, \quad \frac{M_3}{1-\dfrac{M_2}{1-M_1}}, \quad \frac{M_4}{1-\dfrac{M_3}{1-\dfrac{M_2}{1-M_1}}}, \cdots$$

are all on the open interval (0, 1).

10. The pinwheel. If the ladder is bent into a loop, elimination of the unknown coefficients c_i is more difficult, because it leads to an equation involving a continued fraction. We shall discuss the special case shown in Fig. 5, in which the ladder has been bent into a loop and one side has been shrunk to a point. With the convention that $c_{m+1} = c_1$, the method of the preceding section leads to

$$M_i < c_i(1-c_{i+1}), \qquad i = 1, 2, \cdots, m, \tag{10}$$

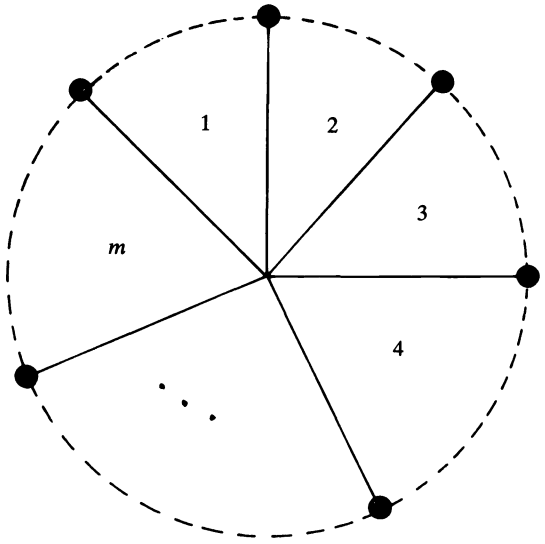


FIG. 5

as a sufficient condition for stable admissibility. The condition is necessary and sufficient if the only strong links are the ones on the circumference. Here $M_i = A_i/(4B_i)$ as before, where A_i is the measure of asymmetry and B_i the measure of strength of the strong link in the ith pie-shaped loop. The sole condition on the constants c_i is $0 < c_i < 1$.

If $m = 2$, a necessary and sufficient condition for existence of c_i is easily shown to be

$$M_1 + M_2 + 2\sqrt{M_1 M_2} < 1.$$

When $m = 3$, a necessary and sufficient condition is

$$M_1 + M_2 + M_3 + 2\sqrt{M_1 M_2 M_3} < 1, \tag{11}$$

as seen next.

When $m = 3$, a brief argument shows that (10) implies the inequality

$$M_1 + M_2 + M_3 < 1 \tag{12}$$

which is needed later. For fixed $c_1 = x$ let us make an optimum choice of c_2 and c_3. As in the discussion of § 9, we should choose c_2 as large as possible, and also c_3 as large as possible. This gives the inequalities

$$M_1 < x < 1, \qquad \frac{M_1}{1 - M_2} < x < 1 \tag{13}$$

where clearly the first is superseded by the second. If these hold we can can make the optimum choice, and if one of these fails, no c_i will do even when $M_3 = 0$. It is seen thus that c_i can be determined if, and only if,

$$M_3 < (1 - x)\left(1 - \frac{M_2 x}{x - M_1}\right) \tag{14}$$

for some x satisfying the inequalities (13).

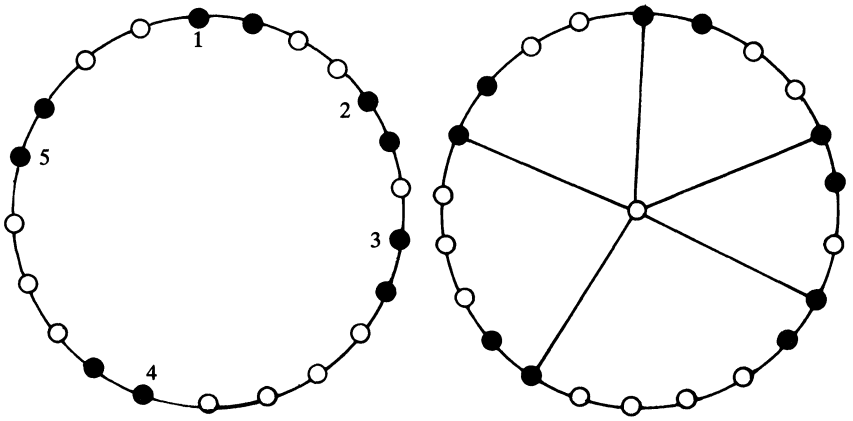
By an elementary calculation, it is found that the maximum occurs at

$$x = M_1 + \left(M_1 M_2 \frac{1-M_1}{1-M_2}\right)^{1/2}$$

This choice of x satisfies (13) and reduces (14) to a relation which, in view of (12), is equivalent to (11). In terms of A_i and B_i the condition is

$$\frac{A_1}{B_1}+\frac{A_2}{B_2}+\frac{A_3}{B_3}+\left(\frac{A_1A_2A_3}{B_1B_2B_3}\right)^{1/2}<4. \tag{15}$$

11. The necklace. Here we consider the effect of having more than one strong link in a single loop. It is assumed that the strong links are *separated*, in the sense that at least one open circle ○ is between any two of them. To identify links, let us label the first black dot of each link as we walk around the loop in the negative (clockwise) direction. If there are m links the labels i run consecutively from 1 to m, and the strength of the ith link is B_i. Separated links with their labels are shown in Fig. 6a for $m = 5$.



(a) (b)

FIG. 6

Let us add a vertex $i = 0$ and corresponding links joining this vertex to the nodes $i = 1, 2, \cdots, m$ introduced above; see Fig. 6b. The new coefficients p_{ij} with $i = 0$ or $j = 0$ are required to satisfy

$$a_0 p_{0i} + a_i p_{i0} = 0, \qquad i = 0, 1, 2, \cdots, m, \tag{16}$$

where a_i are the positive constants associated with the rest of the p_{ij} and where $a_0 > 0$ is arbitrary. The usefulness of this construction depends on three properties described next.

(i) The new terms in $w_0 w_i$ added to the quadratic form all have coefficient 0, hence the new form agrees in value with the old one. This shows that admissibility and stable admissibility hold for the one case if for the other.

(ii) Let R_i and A_i be the ratio and corresponding measure of asymmetry for the small pie-shaped loop containing the ith link in Fig. 6b, where it is understood that these links are traversed in the negative direction. In computing R_i and R_{i+1} the radial path to vertex i (the spoke of the wheel) is traversed twice, once in one direction and once in the other. Hence, the factors for this part of the path cancel out in the product and we

conclude that

$$R_1R_2\cdots R_m = R \tag{17}$$

where R is the corresponding ratio for the loop in Fig. 6a. This calculation involves considerations similar to those in §§ 6 and 7.

(iii) Since the coefficients p_{ij} for $i=0$ or $j=0$ can be chosen at will, subject to (16), there is no restriction on the R_i other than (17). In other words, if R_i are any positive quantities whose product is R, we can choose the coefficients in such a way that the R_i in Fig. 6b agree with these.

We now apply Theorem 1 to Fig. 6b, noting that $n(i)=1$. The result is that the graph of Fig. 6b, and hence also of Fig. 6a, is stably admissible if $A_i<4B_i$ for $i=1,2,\cdots,m$, where

$$A_i = R_i+\frac{1}{R_i}-2.$$

Since R and $1/R$ are interchangeable in our analysis, there is no loss of generality in orienting the outer loop so that $R>1$ and in choosing the new coefficients in such a way that also each $R_i>1$. Then $A_i<R_i-1$ and a sufficient condition is $R_i<1+4B_i$. Since the sole restriction on the R_i is that their product shall be R, we conclude finally that a sufficient condition for stable admissibility is

$$1\leqq R\leqq(1+4B_1)(1+4B_2)\cdots(1+4B_m). \tag{18}$$

It follows from (20) below that one or more of the terms $4B_i$ could be replaced by $2\sqrt{B_i}$, which is an improvement whenever $B_i<1/4$.

On the other hand since $A_i>R_i-2$, a necessary condition is

$$R<(2+4B_1)(2+4B_2)\cdots(2+4B_m). \tag{19}$$

The gap between (18) and (19) can be filled by solving for R_i in terms of A_i, when $R_i>1$, and noting that the resulting relationship is monotone. This gives the necessary and sufficient condition

$$1\leqq R<\prod_{i=1}^{m}[1+2B_i+2(B_i+B_i^2)^{1/2}]. \tag{20}$$

If the B_i are numerous and large, these results represent a vast improvement on the result obtained by breaking a single link as in the proof of Theorem 1.

12. The necklace, continued. Let us now drop the hypothesis that the strong links are separated. This means that a black dot can be common to two strong links. In such a case the corresponding coefficient p_{ii} must be replaced by

$$(1-c_i)p_{ii}+c_ip_{ii}, \qquad 0<c_i<1,$$

as in §§ 9 and 10. If the *successive groups* of black dots are separated, the optimum determination of c_i proceeds much as in § 9, and if the black dots form a *loop* (that is, if every vertex is black) the problem reduces to that in § 10. In particular, (15) suffices when $m=3$. However, there is a further complication, because one must still determine an optimum decomposition $R=R_1R_2\cdots R_m$. A brief argument, which we omit, shows that the optimum choice of c_i is close to the value $c_i=\frac{1}{2}$ used in Theorem 1, provided each R_i is large.

If $c_i=\frac{1}{2}$ (which is a permissible choice whether optimum of not), it is easy to extend the results of § 11 to the case under consideration here. With the same consecutive

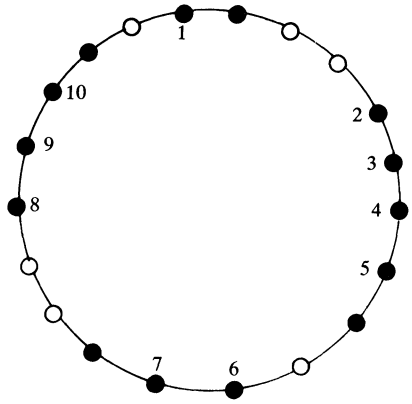


FIG. 7

numbering of strong links as before, let $n(i) = 1$ if the ith link is isolated, $n(i) = 2$ if the ith link has exactly one black dot adjacent to it, and $n(i) = 4$ if the ith link has two black dots adjacent to it. For example in Fig. 7 we have

$$n(i) = 1, 2, 4, 4, 2, 2, 2, 2, 4, 2$$

for $n = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$, respectively. Then the results of § 10 remain valid without the separation condition, provided each B_i is replaced by $B_i/n(i)$.

In particular, if every vertex of the loop has a black dot, then $n(i) = 4$ and the basic inequalities become $A_i < B_i$ instead of $A_i < 4B_i$. If all the B_i are equal in this case we may as well take all the A_i to be equal too, hence $R_i = R^{1/m}$ where m, the number of strong links, is also the number of nodes. Thus a sufficient condition for stable admissibility is

$$R^{1/m} + R^{-1/m} - 2 < B,$$

where $B_i = B$ is the common measure of strength.

Applying these results to each loop in a general graph, we get a sharpened version of Theorem 1 in which extra strong links are not merely ignored, but are exploited. We do not give a formal statement because the same result can be achieved by adding fictitious nodes and branches as in Fig. 6b and applying Theorem 1 as it stands to the new graph so obtained. The fictitious nodes are introduced only for loops that have more than one strong link.

Acknowledgment. The authors are grateful to the referee for his careful reading of this paper and for his constructive suggestions.

REFERENCES

[1] F. HARARY AND E. M. PALMER, *Graphical Enumeration*, Academic Press, New York, 1973.
[2] N. KRIKORIAN, *The Volterra model for three species predator-prey systems: Boundedness and stability*, J. Math. Biol., 7 (1979), pp. 117–132.
[3] J. S. MAYBEE, *Combinatorially symmetric matrices*, Linear Algebra Appl., 8 (1974), pp. 529–537.
[4] J. MAYBEE AND J. QUIRK, *Qualitative problems in matrix theory*, SIAM Rev., 2 (1969), pp. 32–51.
[5] S. Y. PARTER AND J. W. T. YOUNGS, *The symmetrization of matrices by diagonal matrices*, J. Math. Anal. Appl., 4 (1962), pp. 102–110.
[6] R. REDHEFFER AND Z. ZHOU, *Global asymptotic stability for a class of many-variable Volterra prey-predator systems*, J. Nonlinear Anal. Theor. Meth. Appl. to appear.
[7] V. VOLTERRA, *Leçons sur la théorie mathématique de la lutte pour la vie*, Gauthier-Villars et Cie., Paris, 1931.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 1, March 1982

0196-5212/82/0301-0013 $01.00/0

AN ODD ORDER SEARCH PROBLEM*

R. S. BOOTH†

Abstract. Select $k+1$ points in a given interval. Successively remove an end subinterval and select a new point in the remaining interval. It is desired to calculate how fast the sequence of lengths of successive intervals decreases. The answer has long been known when k is even or when $k=1$ or 3. The present paper deals with $k=5$ and exhibits the difficulties in proceeding to higher odd k.

1. Introduction. Let $D=[a, b]$ be a compact interval of the real line, and let k be a positive integer. Select $(k+1)$ points $x_1, x_2, \cdots, x_{k+1}$ in D so that

$$a<x_{k+1}<x_k<\cdots<x_1<b.$$

This divides D into $k+2$ intervals. The procedure is to remove either of the two end subintervals, select one new point in the remainder and successively repeat until a total of n points (including the original $k+1$) have been selected. At this stage, after removing an end subinterval, a final interval $D'\subset D$ is produced. n is a preassigned integer. The rule for selecting each point is called a strategy. The aim is to describe a strategy which minimizes the maximum length of all possible final intervals D' obtained this way.

More particularly, suppose D has length d, and let S be any particular strategy. Let $d_n(S, k)$ be the maximum length of all final intervals obtained by employing S with n points, and put

$$\rho(S, k)=\lim_{n\to\infty}\inf\,[d/d_n(S, k)]^{1/n}, \qquad \rho(k)=\sup_S \rho(S, k).$$

$\rho(S, k)$ is a measure of the efficiency of the strategy. It is the "average" ratio of lengths of intervals of consecutive steps in the strategy.

The problem as stated above arises when one attempts to locate the zero of the kth order derivative of a suitable function from the class of all functions defined and possessing a continuous kth order derivative on D. The simplest such problem, when $k=0$, leads to the classical bisection strategy S, for which $\rho(S, 0)=2$. Indeed, $\rho(0)=2$. J. Keifer [3], seems to have been the first author to examine this problem for $k\geqq 1$. When $k=1$, we have the classic Fibonacci search problem for locating maxima or minima. We refer the reader to Kiefer's paper for this and other background information. Also relevant is some of the material in the book *Applied Dynamic Programming* by Bellman and Dreyfus [1, Chap. 4]. The general problem was examined by C. L. Mallows (unpublished, to the author's knowledge). Mallows produced an efficient strategy, S, valid for all even k. It is efficient in the sense that $\rho(S, k)=\rho(k)$ for all even k, and moreover, $\rho(k)=2^{2/(k+2)}$ for all even k. This result is described by the author [2].

It is known [2] that, if k is odd, if $p=(k+1)/2$ and if β_p is the positive root of the equation

$$x^{p+1}=x+1,$$

then $\rho(k)=\beta_p$ when $k=1$ and $k=3$.

* Received by the editors March 18, 1980, and in final revised form June 4, 1981.

† School of Mathematical Sciences, Flinders University of South Australia, Bedford Park, South Australia 5042.

Moreover,

$$2^{2/(k+3)} \leqq \rho(k) \leqq \beta_p < 2^{2/(k+1)}$$

for all odd k.

It is natural, on the basis of the above, to conjecture that $\rho(k) = \beta_p$ for all odd k. The conjecture is an attractive one, especially after the simplicity of the even k and $k = 1$, $k = 3$. The present paper shows, however, that the conjecture is false for $k = 5$, suggesting that, in particular, $\rho(3)$ equals β_2, as derived in [2], only because of a "lucky" circumstance.

Henceforth, $k = 5$ and, thus $k + 1 = 6$. An interval together with its six selected points will be called a configuration. If the interval is $[a, b]$ and the six points are x_1, $x_2, \cdots, x_6$, with $x_1 > x_2 > \cdots > x_6$, then this will be written

$$[a;\ x_6, x_5, x_4, x_3, x_2, x_1;\ b].$$

The procedure leads to either the configuration

$$[a;\ x_6, x_5, x_4, x_3, x_2;\ x_1],$$

in which one new point $\mathbf{y}$ must be selected, or to

$$[x_6;\ x_5, x_4, \cdots, x_1;\ b],$$

in which one new point $\mathbf{y}'$ must be selected. The new point is inserted in its correct position with regard to order, and is written in boldface.

It is permitted, if desired, to choose two (or more) new points after the removal of an end subinterval. In this case, one (or more) of the x_i must be removed from the configuration, and ignored, so that the new configuration still contains exactly six points in its interior. Removal is indicated by a caret $\hat{}$.

2. The result.

THEOREM. *Let λ be the positive root of the equation*

$$\lambda^{12} = 2\lambda^8 + 1. \tag{1}$$

Then

$$\rho(5) = \lambda.$$

Remark.

$$\lambda = 1.2186533 \cdots < \beta_3 = 1.2206 \cdots.$$

Proof. We show first that $\rho(5) \leqq \lambda$. Let

$$L_n = \sup\{L > 0\text{: there is a strategy } S, \text{ such that if } D = [0, L], \text{ then } d_n(S, k) \leqq 1\}.$$

Commence with the interval $[0, L_n]$, containing six initial points $x_1, x_2, \cdots, x_6$, such that

$$0 < x_6 < x_5 < \cdots < x_1 < L_n.$$

This is the configuration

$$[0;\ x_6, x_5, x_4, x_3, x_2, x_1;\ L_n].$$

As in [2], it is necessary that

$$x_3 \leqq L_{n-3}, \qquad L_n - x_3 \leqq L_{n-4}, \tag{2}$$

$$x_4 \leqq L_{n-4}, \qquad L_n - x_4 \leqq L_{n-3}. \tag{3}$$

After three stages of reduction, we may be left with the interval $[0, x_3]$, containing x_4, x_5, x_6 and three other points. It is easy to see that these three points must be less than x_4. So, by relabelling x_5 and x_6 if necessary, we have the configuration

$$[0;\ x_9, x_8, x_7, x_6, x_5, x_4;\ x_3],$$

in which, in particular, $x_6 \leqq L_{n-6}$.

After four further stages, we may be restricted to the interval $[x_6, x_3]$, which must contain x_4, x_5 and four other points. There are two possibilities. Either at least two of the new points are greater than x_4, in which case

$$x_4 - x_6 \leqq L_{n-10}, \tag{4}$$

or at least three of the new points are less than x_4, in which case

$$x_3 - x_4 \leqq L_{n-12}. \tag{5}$$

It follows by (2) and (3) that either $L_n - L_{n-3} - L_{n-6} \leqq L_{n-10}$ or $L_n - L_{n-4} - L_{n-4} \leqq L_{n-12}$. Hence,

$$L_n \leqq \max\begin{cases} 2L_{n-4} + L_{n-12}, \\ L_{n-3} + L_{n-6} + L_{n-10}, \end{cases} \tag{6}$$

at least for $n \geqq 13$.

$L_n = 1$ for $1 \leqq n \leqq 5$, and $L_n \leqq \lambda^n$ for $6 \leqq n \leqq 13$. It is easy to show by induction on n that $L_n \leqq \lambda^n$ for all n, by using (6). (The fact that $\lambda^{-3} + \lambda^{-6} + \lambda^{-10} < 1$ is needed.) Hence, $\rho(5) \leqq \limsup_{n\to\infty} L_n^{1/n} \leqq \lambda$, which completes the first part of the proof.

In order to establish the reverse inequality, we must describe a strategy S for $\rho(S, 5) \geqq \lambda$. It is possible to employ a sequence $\{U_n\}$ such that $U_n \geqq L_n$ for all n, where L_n is as defined earlier, with $U_n = 2U_{n-4} + U_{n-12}$ at least for $n \geqq 14$. We prefer to use an asymptotic strategy, which is somewhat easier to describe.

We introduce the notation

$$\begin{aligned} t &= \lambda^{-4} && (=0.453397\cdots), \\ \mu &= t/(1-t) && (=0.829484\cdots), \\ x &= 1 - t^3\ (=2t) && (=0.906794\cdots), \\ y &= 1 - \mu^3 t^3 && (=0.946806\cdots). \end{aligned} \tag{7}$$

Notice that

$$1 = 2t + t^3, \tag{8}$$

$$1 - \mu = \mu t^2, \tag{9}$$

$$1 - \mu^2 = \mu^2 t, \tag{10}$$

$$1 - \mu^3 = ty. \tag{11}$$

Define $\{d_n\}$, $n = 0, 1, 2, \cdots$, where, if $n = 4m + q$, m integer, $q = 0, 1, 2$ or 3, then

$$d_n = \mu^q t^m.$$

The strategy S will describe how to proceed from a configuration on an interval of length d_n, to one of d_{n+p}, for some positive integer p. This yields

$$\rho(S, 5) = \liminf_{n\to\infty} [1/\mu^q t^m]^{1/n} = 1/t^{1/4} = \lambda$$

This implies $\rho(5) \geqq \lambda$ and hence, establishes the theorem.

We need just one lemma, the result of which is easy to see.

LEMMA. *Given the configuration*

$$[x_7;\ x_6, x_5, x_4, x_3, x_2, x_1;\ x_0],$$

suppose there is a strategy for which, after $n-6$ *further points are selected, the final interval has length at most* d' *for some* $d' < x_0 - x_7$. *If*

$$[y_7;\ y_6, y_5, \cdots, y_1;\ y_0]$$

is another configuration, in which

$$y_i - y_{i+1} \leqq x_i - x_{i+1}$$

for

$$i = 0, 1, 2, 3, 4, 5, 6,$$

then there is a strategy for which, after $n-6$ *further points are selected, the final interval for this second configuration has length at most* d'.

We return now to the theorem itself, and the construction of the strategy. We begin with the configuration

$$A_1 = [0;\ \mu^2 t, \mu t, t, \mu^3, \mu^2, \mu;\ 1].$$

(There is no loss of generality in scaling the initial interval to one of length 1.)

If the right-hand subinterval at A_1 is removed, we select the new point $\mu^3 t$, indicated in boldface, to obtain

$$B_1 = [0;\ \boldsymbol{\mu^3 t}, \mu^2 t, \mu t, t, \mu^3, \mu^2;\ \mu].$$

We identify B_1 with the configuration A_μ below, and consider it later on. If, instead, the left-hand subinterval of A_1 is removed, we must select a new point in the interval $[\mu^2 t, 1]$. It is convenient (and unless otherwise indicated, we will automatically do so when the left-end subinterval is removed) to transfer the origin to the right-end point, in this case 1, and reverse the orientation. This gives

$$[0;\ 1-\mu, 1-\mu^2, 1-\mu^3, 1-t, 1-\mu t;\ 1-\mu^2 t].$$

In this new coordinate system, we simplify the numbers $1-\mu$, $1-\mu^2$, etc., by using (7)–(11); select the new point $\boldsymbol{\mu t}$, again indicated by boldface, to get

$$B_2 = [0;\ \mu t^2, \mu^2 t, \boldsymbol{\mu t}, ty, 1-t, 1-\mu t;\ \mu^2].$$

Now we proceed similarly from B_2, obtaining

$$C_1 = [0;\ \mu t^2, \boldsymbol{\mu^3 t}, \mu^2 t, \mu t, ty, 1-t;\ 1-\mu t],$$

which applies if the right-hand end of B_2 is removed. Similarly, we obtain

$$C_2 = [0;\ \mu^2 t^3, \mu^2 t^2, \mathbf{t^2}, \mu^3 t, \mu^2 t, \mu t;\ \mu^3 x]$$

if the left subinterval of B_2 is removed. As before, the origin is shifted and the orientation is reversed. Now C_1 leads to both

$$D_1 = [0;\ \mu t^2, \mathbf{t^2}, \mu^3 t, \mu^2 t, \mu t, ty;\ \mu^3 - \mu^3 t^4],$$

$$D_2^T = [0;\ \widehat{t^3 y}, \mu^2 t^2, \boldsymbol{\mu t^2}, t^2, \mu^3 t, \mathbf{\mu^2 t}, \mu t;\ t],$$

and C_2 leads to both

$$D_3 = [0;\ \widehat{\mu^2 t^3}, \boldsymbol{\mu^3 t^2}, \mu^2 t^2, \boldsymbol{\mu t^2}, t^2, \mu^3 t, \mu^2 t;\ \mu t],$$

$$D_4 = [0;\ \mu^2 t^2, \boldsymbol{\mu t^2}, t^2, \mu^3 t, \mu^2 t, \mu t;\ t].$$

The superscript T for the configuration D_2 denotes a departure form the convention regarding choice of origin and orientation. The orientation is reversed from what it would be if the T were not present. Thus, D_2^T is obtained from C_1 by removing $[0, \mu t^2]$ from C_1 and subtracting μt^2 from all points in the remainder of C_1. This is done for ease of identification.

The configuration D_4 is precisely the configuration A_1 scaled down by the factor $t = \lambda^{-4}$. At most four new points have been selected to arrive at D_4. So we may identify D_4 with A_1, in the obvious sense that the strategy which applied to A_1 also applies to D_4.

The same remark applies to D_2^T, where we selected two new points simultaneously as indicated and ignored one other point, namely t^3y, to produce the required identification. The caret is used to denote the removal of an unwanted point.

D_1 is identified with the configuration A_{μ^3} considered later on. This involves an application of the lemma, which applies trivially with the choice

$$A_{\mu^3} = [0;\ \mu t^2, t^2, \mu^3 t, \mu^2 t, \mu t, ty;\ \mu^3].$$

We observe that only three new points have been selected to arrive at D_1, of length at most μ^3.

Finally, with t as scaling factor, D_3 can be identified with A_μ, which we consider next.

The procedure and pattern for handling A_μ, and later on A_{μ^2}, A_{μ^3}, A_1^m and $A_{\mu^3}^{m_3}$, is the same as that for A_1. For this reason, we adopt the same notation and conventions as previously, including the labelling of the configurations. No confusion need arise if all configurations B, C, D and E are taken in context.

In A_μ, for reasons of future identification, it is convenient to allow for either of two possible locations for the second point from the right. We take

$$A_\mu = [0;\ \mu^3 t, \mu^2 t, \mu t, t, \mu^3 x \text{ or } \mu^3, \mu^2;\ \mu].$$

This leads to

$$B_1 = [0;\ \mathbf{t}^2, \mu^3 t, \mu^2 t, \mu t, t, \mu^3 x \text{ or } \mu^3;\ \mu^2]$$

and

$$B_2 = [0; \mu^2 t^2, \boldsymbol{\mu^3 t}, \boldsymbol{\mu^2 t}, \mu t, t, \mu^3 x; \mu^3].$$

Notice that, since $\mu - \mu^3 = \mu^3 t$ and $\mu - \mu^3 x = \mu^2 t$, one of the two "new" points in B_2 is already present, so we need only to select the other.

B_1 is identical to A_{μ^2} considered below. Proceeding from B_2, we have

$$C_1 = [0;\ \mu^2 t^2, \mathbf{t}^2, \mu^3 t, \mu^2 t, \mu t, t;\ \mu^3 x],$$
$$C_2 = [0; \mu^3 t^3, \mu^3 t^2, \boldsymbol{\mu t^2}, t^2 y, \mu^3 t, \mu^2 t; t - \mu^3 t^4]$$

and hence, as before,

$$D_1 = [0;\ \mu^2 t^2, \boldsymbol{\mu t^2}, t^2, \mu^3 t, \mu^2 t, \mu t;\ t],$$
$$D_2 = [0;\ \widehat{\mu^2 t^3}, \boldsymbol{\mu^3 t^2}, \mu^2 t^2, \boldsymbol{\mu t^2}, t^2, \mu^3 t, \mu^2 t;\ \mu t],$$
$$D_3 = [0;\ \widehat{\mu^3 t^3}, \mathbf{t}^3, \mu^3 t^2, \boldsymbol{\mu^2 t^2}, \mu t^2, t^2 y, \mu^3 t;\ \mu^2 t],$$
$$D_4 = [0;\ \mu^3 t^2, \boldsymbol{\mu^2 t^2}, \mu t^2, \mathbf{t}^2, \mu^3 t x, \widehat{\mu^3 t}, \mu^2 t;\ \mu t].$$

These can be identified respectively with A_1, $\bar{A}_\mu$, A_{μ^2}, A_μ, all scaled by the factor t.

Three new points were selected passing from A_μ to D_1, four new ones in the case of D_2, D_3, D_4.

With the same notation as before, we consider A_{μ^2} with two alternative independent possibilities. We put

$$A_{\mu^2} = [0;\ t^2, \mu^3 t, \mu^2 t, \mu t, ty \text{ or } t, \mu^3 x \text{ or } \mu^3;\ \mu^2].$$

This leads to

$$B_1 = [0;\ \boldsymbol{\mu t}^2, t^2, \mu^3 t, \mu^2 t, \mu t, t \text{ or } ty;\ \mu^3 x \text{ or } \mu^3],$$

which we identify with A_{μ^3} (at the right end we can replace $\mu^3 x$ with μ^3 if desired) and

$$B_2 = [0;\ \boldsymbol{\mu^3 t}^2, \boldsymbol{\mu t}^2, \mu^3 tx \text{ or } \mu^3 t, \mu^2 t, \mu t, ty;\ \mu^2 - t^2].$$

Again, only one new point needs selection in B_2. Proceeding from B_2, we obtain first

$$C_1 = [0;\ \mu^3 t^2, \mu t^2, \mathbf{t}^2, \mu^3 tx \text{ or } \mu^3 t, \mu^2 t, \mu t;\ ty],$$

$$C_2 = [0;\ \mu^3 t^3, \mu^3 t^2 x, \mu t^2, \mathbf{t}^2\mathbf{y}, \mu^3 tx - \mu^3 t^5 \text{ or } \mu^3 t - \mu^3 t^5, \mu^2 t;\ \mu t - \mu^3 t^5],$$

Hence, with the same pattern as before,

$$D_1 = [0;\ \mu^3 t^2, \boldsymbol{\mu^2 t}^2, \mu t^2, t^2, \mu^3 tx \text{ or } \mu^3 t, \mu^2 t;\ \mu t],$$

$$P = D_2 = [0;\ \mu^3 t^3, \mu^3 t^2, \boldsymbol{\mu t}^2, \mathbf{t}^2\mathbf{y}, \mu^3 t - \mu t^4, \mu^3 t;\ \mu^2 t],$$

$$D_3^T = [0;\ \widehat{\mu^2 t^3 \text{ or } t^3 y}, \mathbf{t}^3, \mu^3 t^2, \mu^2 t^2, \boldsymbol{\mu t}^2, t^2, \mu^3 t;\ \mu^2 t],$$

$$D_4^T = [0;\ \widehat{\mu^3 t^3}, \mathbf{t}^3, \mu^3 t^2, \mu^2 t^2, \boldsymbol{\mu t}^2, \mathbf{t}^2\mathbf{y}, \mu^3 t;\ \mu^2 t].$$

Here D_1 is A_u and D_3, D_4 are A_{μ^2}, all scaled by the factor t. D_2 here needs separate identification; we name it P and deal with it later. In D_2, only one of μt^2, $t^2 y$ needs selection, likewise in D_4.

Of the "cycle" A_1, A_μ, A_{μ^2}, A_{μ^3} of configurations, only A_{μ^3} is left. We begin with

$$A_{\mu^3} = [0;\ \mu t^2, t^2, \mu^3 t, \mu^2 t, \mu t, yt \text{ or } t;\ \mu^3].$$

It leads to

$$B_1 = [0;\ \boldsymbol{\mu^2 t}^2, \mu t^2, t^2, \mu^3 t, \mu^2 t, \mu t;\ t],$$

which is A_1 scaled by the factor t, and

$$B_2 = [0;\ \mu^3 t^2 \text{ or } \mu^2 t^2, \boldsymbol{\mu t}^2, t^2 y, \mu^3 t, \mu^2 t, \mu t - \mu^3 t^5;\ t - \mu^3 t^3].$$

From B_2 follows

$$C_1 = [0;\ \boldsymbol{\mu^3 t}^2, \boldsymbol{\mu^2 t}^2, \mu t^2, t^2 y, \mu^3 t, \mu^2 t;\ \mu t - \mu^3 t^5],$$

$$C_2 = [0;\ \mu t^4, t^3 y, \boldsymbol{\mu^3 t}^2 - \boldsymbol{\mu^3 t}^6, \mu^2 t^2, t^2, \mu^3 t - \mu^2 t^4;\ \mu^2 tx]$$

and, as before

$$D_1 = [0;\ \mathbf{t}^3, \mu^3 t^2, \mu^2 t^2, \mu t^2, t^2 y, \mu^3 t;\ \mu^2 t],$$

$$Q = D_2 = [0;\ \mu^3 t^3, \mu^3 t^2 x, \boldsymbol{\mu^2 t}^2, \mu t^2, t^2 y, \mu^3 t - \mu t^4;\ \mu^3 t - \mu^3 t^5],$$

$$R = D_3 = [0;\ \mu t^4, t^3 y, \mu^3 t^2 - \mu^3 t^6, \boldsymbol{\mu^2 t}^2\mathbf{x}, \mu^2 t^2, t^2;\ \mu^3 t - \mu^2 t^4],$$

$$D_4 = [0;\ \widehat{\mu^3 t^3}, \mu t^3, \mathbf{t}^3, \boldsymbol{\mu^3 t}^2, \mu^2 t^2, \mu t^2, t^2 y;\ \mu^3 t - \mu^3 t^5].$$

Of these, D_1 identifies with A_{μ^2}, D_4 identifies with A_{μ^3}, by the scale factor t and application of the lemma.

Of course, at this point, the construction would be complete if we knew what to do with the three outstanding configurations P, Q and R. After some experimentation, we are led to consider the following modifications to A_1 and A_{μ^3}, which we denote respectively by A_1^m and $A_{\mu^3}^{m_3}$. First

$$A_1^m = [0;\ \mu^3 t, \mu t, t, \mu^3 x \text{ or } \mu^3, \mu^2, \mu;\ 1-\mu^3 t^3].$$

It is not hard to show that if the right-hand end is 1, then the choice $\mu^3 x$ as indicated would not be available. We proceed as before, obtaining successively

$$B_1 = [0;\ \mu^3 t, \boldsymbol{\mu^2 t}, \mu t, t, \mu^3 x \text{ or } \mu^3, \mu^2;\ \mu],$$

which is A_μ,

$$B_2 = [0;\ \mu^3 t^2, \mu^3 t, \boldsymbol{\mu t}, \mathbf{t}y, \mu^3 - \mu t^3, \mu^3;\ \mu^2],$$
$$C_1 = [0;\ \mu^3 t^2, \mu^3 t, \boldsymbol{\mu^2 t}, \mu t, ty, \mu^3 - \mu t^3;\ \mu^3],$$
$$C_2 = [0;\ \mu^3 t^2, t^2 y, \mu^3 t, \mu^2 t, \boldsymbol{\mu t}, ty;\ \mu^3]$$

and

$$D_1 = [0;\ \mu^3 t^2, \mathbf{t}^2, \mu^3 t, \mu^2 t, \mu t, ty;\ \mu^3 - \mu t^3],$$
$$D_2^T = [0;\ \mu^2 t^2, \boldsymbol{\mu t^2}, \widehat{t^2 y}, \mathbf{t}^2, \mu^3 t, \mu^2 t, \mu t;\ t],$$
$$D_3 = [0;\ \mu^3 t^2, \boldsymbol{\mu t^2}, \widehat{t^2 y}, \mathbf{t}^2, \mu^3 t, \mu^2 t, \mu t;\ ty],$$
$$D_4 = [0;\ \mu^2 t^2, \boldsymbol{\mu t^2}, \widehat{t^2 y}, \mathbf{t}^2, \mu^3 t, \mu^2 t, \mu t;\ t].$$

Of these last four, D_2^T and D_4 are suitably scaled copies of A_1, and D_3 is a suitably scaled copy of A_1^m. Moreover, the configuration P is now not outstanding; it is B_2 above, scaled by the factor t. D_1 is identified with $A_{\mu^3}^{m_3}$ below.

We put

$$A_{\mu^3}^{m_3} = [0;\ \mu^3 t^2, t^2, \mu^3 tx \text{ or } \mu^3 t, \mu^2 t, \mu t, ty;\ \mu^3 - \mu t^3].$$

This leads to

$$B_1 = [0;\ \mu^3 t^2, \boldsymbol{\mu t^2}, t^2, \mu^3 tx \text{ or } \mu^3 t, \mu^2 t, \mu t;\ ty],$$

which is A_1^m scaled by t, and

$$B_2 = [0;\ \mu^2 t^3, \mu^3 t^2, t^2 - \mu^3 t^4, \mathbf{t}^2, \mu^3 tx \text{ or } \mu^3 t, \mu^2 t - \mu^3 t^4;\ \mu t],$$
$$C_1 = [0;\ \mu^2 t^3, \mu^3 t^2, \boldsymbol{\mu^2 t^2}, t^2 - \mu^3 t^4, t^2, \mu^3 tx \text{ or } \mu^3 t;\ \mu^2 t - \mu^3 t^4],$$
$$C_2 = [0;\ \widehat{t^3 y}, \mathbf{t}^3, \boldsymbol{\mu^3 t^2}, \boldsymbol{\mu^2 t^2}, \mu t^2, t^2 y\, \mu^3 t;\ \mu^2 t].$$

C_2 is A_{μ^2} scaled by t; we need only proceed from C_1, which leads to

$$D_1 = [0;\ \mu^2 t^3, \mu^3 t^2, \mu^2 t^2, \boldsymbol{\mu t^2}, t^2 - \mu^3 t^4, t^2;\ \mu^3 t],$$
$$D_2^T = [0;\ \mu^3 t^3, \widehat{\mu t^3}, \mathbf{t}^3, \mu^3 t^2, \mu^2 t^2, \boldsymbol{\mu t^2}, \mathbf{t}^2 y;\ \mu^3 t - \mu t^4].$$

Since D_2^T is identifiable with $A_{\mu^3}^{m_3}$, scaled by t, we proceed from D_1 only, to obtain

$$E_1 = [0;\ \mu^2 t^3, \boldsymbol{\mu t^3}, \mathbf{t}^3, \mu^3 t^2, \mu^2 t^2, \mu t^2, \widehat{t^2 - \mu^3 t^4};\ t^2],$$

which is A_1 scaled by t^2, and

$$E_2 = [0;\ \mu^3 t^3, \mu t^3, \widehat{t^3 y}, \mathbf{t}^3, \mu^3 t^2, \mu^2 t^2, \boldsymbol{\mu t^2};\ t^2 y],$$

which is A_1^m scaled by t_2.

It is now easy to resolve the configurations Q and R. Consider Q, which is

$$Q=[0;\ \mu^3t^3, \mu^3t^2x, \mu^2t^2, \mu t^2, t^2y, \mu^3t-\mu t^4;\ \mu^3t-\mu^3t^5].$$

Removing its right end subinterval, we obtain

$$E_1=[0;\ \mu^3t^3, \mathbf{t}^3, \mu^3t^2x, \mu^2t^2, \mu t^2, t^2y;\ \mu^3t-\mu t^4],$$

which is $A_{\mu^3}^{m_3}$ scaled by the factor t and, hence, can be treated like $A_{\mu^3}^{m_3}$, as above. On the other hand, if the left subinterval of Q is removed, we obtain, with simplification and the usual convention of orientation,

$$E_2=[0;\ \widehat{\mu^3t^4}, \mu^3t^3,\ \mu t^3, \mathbf{t}^3, \mu^3t^2x, \mu^2t^2, \boldsymbol{\mu}\mathbf{t}^2;\ t^2y].$$

Since this is A_1^m scaled by t^2, we need pursue this no further.

Finally, since

$$R=[0;\ \mu t^4, t^3y, \mu^3t^2-\mu^3t^6, \mu^2t^2x, \mu^2t^2, t^2;\ \mu^3t-\mu^2t^4],$$

we obtain, after removing the right subinterval and then reversing the orientation,

$$E_1^T=[0;\ \mu^2t^3, \mu t^3, t^3, \mu^3t^2, \boldsymbol{\mu}^2\mathbf{t}^2, \mu t^2;\ t^2],$$

which is A_1 scaled by the factor t_2. If, instead, the left end subinterval is removed, we preserve the orientation by subtracting μt^4 from all terms, to obtain

$$E_2^T=[0;\ \mu^3t^3, \mu t^3, t^3, \mu^3t^2x, \boldsymbol{\mu}^2\mathbf{t}^2, \mu t^2;\ t^2y].$$

This is A_1^m scaled by t^2, considered earlier.

This completes the construction, and the proof of the theorem.

Remark. The above construction was inferred from a study of the sequence $\{U_n\}$, mentioned earlier, defined by

$$U_1=U_2=U_3=U_4=U_5=1,$$
$$U_6=U_7=U_8=2, \qquad U_9=U_{10}=3,$$
$$U_{11}=4, \quad U_{12}=5, \quad U_{13}=6,$$
$$U_n=2U_{n-4}+U_{n-12} \quad \text{for } n\geqq 14.$$

Here, if $n\equiv 2 \pmod 4$,

$$\lim\frac{U_{n-1}}{U_n}=\lim\frac{U_{n-2}}{U_{n-1}}=\lim\frac{U_{n-3}}{U_{n-2}}=\mu,$$

while $\lim U_{n-4}/U_n=t$.

3. Conclusion. It would be pleasant to be able to find at least an analogue of the first part of the theorem, valid for $\rho(k)$, k odd, $k\geqq 7$. The author can show, by a method similar to that used here, that

$$\rho(7)\leqq\lambda=1.166287<\beta_4,$$

where

$$\lambda^{17}=2\lambda^{12}+1,$$

and

$$\rho(9)<\lambda=1.134544<\beta_5,$$

where

$$\lambda^{22} = 2\lambda^{16} + 1.$$

However, the method does not seem to extend easily to higher k, and it is by no means clear whether these inequalities are best possible. One of the main problems is how to compare the relative sizes of the zeros of two polynomials of high degrees.

There is another complication. For $k = 5$, the speed λ arises from the sequence $L_n = 2L_{n-4} + L_{n-12}$, in which the value of $L_n - L_{n-1}$ can be selected by suitable choice of initial conditions to equal L_{n-9} or L_{n-8}. No such choice can be made for the sequence $L_n = 2L_{n-5} + L_{n-17}$ which pertains to the best known estimate for $\rho(7)$.

Acknowledgment. The author wishes to acknowledge the very helpful comments of the referee during the preparation of this paper.

REFERENCES

1. R. BELLMAN AND S. DREYFUS, *Applied Dynamic Programming*, Princeton University Press, Princeton, NJ, 1962.
2. R. S. BOOTH, *Location of zeros of derivatives*, SIAM J. Appl. Math., 15 (1967), pp. 1496–1501.
3. J. KIEFER, *Optimum sequential search and approximation methods under minimum regularity assumptions*, J. Soc. Indust. Appl. Math., (1957), pp. 105–136.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 2, June 1982

0196-5212/82/0302-0001 $01.00/0

COMPUTING AN OPTIMAL INVARIANT CAPITAL STOCK*

PHILIP C. JONES†

Abstract. This paper derives a new termination criterion for Lemke's linear complementarity algorithm which is applied to the problem of computing a capital stock invariant under optimization for an economy with a linear technology and piecewise linear utility. We derive a condition which guarantees that the computed solution is nontrivial. When the economy has a Leontief technology and linear utility, we give a method for solving the associated linear complementarity problem in $O(n^3)$ operations.

1. Introduction. The problem of computing a capital stock invariant under optimization has been studied by several authors. Hansen and Koopmans [5] showed that a fixed point algorithm yields approximate solutions when the utility function is continuously differentiable and increasing. Dantzig and Manne [2] investigated the case of piecewise linear utility and found that Lemke's linear complementarity algorithm can be applied to find a solution.

This paper derives a new termination criterion for Lemke's algorithm which generalizes that obtained by Dantzig and Manne [2]. We also discuss a condition which guarantees nontriviality of the computed solution. Finally, we examine the special case where the economy has a Leontief technology and linear utility. Although the linear complementarity problem, in general, has been shown by Chung [1] to be NP-complete, we find that the linear complementarity problem associated with this special case can be solved in $O(n^3)$ operations.

Section 2 describes the model and shows that solving an associated linear complementarity problem yields a solution. In § 3, we derive the new termination result for Lemke's algorithm and discuss its application to our problem. Section 4 discusses the special case.

2. The economic model. We suppose that the economy has two types of goods. Those that can be produced by various activities are called producible, while those that cannot be produced are called primary and are assumed to be made available at fixed levels in each time period. The technology is then given by $A \in R^{m\times n}$, $B \in R_+^{m\times n}$ and $b \in R^m$, where

A_{ij} denotes the amount of good i used to operate activity j at unit level;
B_{ij} denotes the amount of good i produced by operating activity j at unit level;
b_i denotes the amount of good i exogenously provided in each time period ($b_i < 0$ denotes a good withdrawn for subsistence).

The utility function $u: R^n \to R^1$ is assumed to be piecewise linear, concave and bounded below. We denote by $x_t \in R^n$, $t = 1, 2, \cdots$, the activity levels in period t and consider the following problem, which we denote P(b_0):

P(b_0) Given $b_0 \in R^n$ find $\{x_t\}_{t=1}^{\infty}$ solving

$$\max \sum_{t=1}^{\infty} \delta^{t-1} u(x_t),$$

$$\text{s.t.} \quad Ax_1 \leqq b_0 + b,$$

$$Ax_t \leqq Bx_{t-1} + b, \qquad t = 2, 3, \cdots,$$

$$x_t \geqq 0, \qquad t = 1, 2, \cdots,$$

* Received by the editors December 8, 1980, and in final form July 8, 1981.

† Department of Industrial Engineering and Management Sciences, Technological Institute, Northwestern University, Evanston, Illinois 60201. This paper contains portions of the author's PhD thesis. The work was supported in part by the National Science Foundation under grant MCS74-21222 A02 with the University of California.

where $\delta \in (0, 1)$ is the discount rate. The problem with which we are concerned can then be stated as:

P Find $x \in R^n$ such that $x_t = x$, $t = 1, 2, \cdots$ solves P(Bx).

Having solved problem P, we will have obtained a capital stock invariant under discounted optimization. This is the natural extension of the notion of an optimal stationary program defined by Gale [4] for the notion of overtaking optimality without discounting.

First we note that since the utility function is piecewise linear and concave we can write it as

$$u(x) = \max z, \qquad \text{s.t.} \quad ze \leqq Vx + d,$$

where e is the vector of all ones and row i of V, together with d_i, constitutes the ith linear "piece" of the utility function. It is then straightforward to show that the problem P(b_0) is equivalent to the following linear problem:

Given $b_0 \in R^n$, find $[(z_t, x_t)]_{t=1}^{\infty}$ solving

$$\begin{aligned} &\max \sum_{t=1}^{\infty} \delta^{t-1} z_t, \\ &\text{s.t.} \quad z_t e \leqq Vx_t + d, \qquad t = 1, 2, \cdots, \\ &\qquad Ax_1 \leqq b_0 + b, \\ &\qquad Ax_t \leqq Bx_{t-1} + b, \qquad t = 2, 3, \cdots, \\ &\qquad x_t \geqq 0, \qquad t = 1, 2, \cdots. \end{aligned}$$

This is equivalent to replacing A, B, b, b_0 and $u(x_t)$ in P(b_0) with

$$\bar{A} = \begin{bmatrix} 0 & A \\ e & -V, \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} 0 & B \\ 0 & 0, \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} b \\ d \end{bmatrix}, \quad \bar{b}_0 = \begin{bmatrix} b_0 \\ 0 \end{bmatrix}, \qquad v = [1, 0].$$

This reduction shows that it is sufficient to deal with the case of linear utility. We shall henceforth assume that $u(x) = vx$, where $v \in R^n$.

To see that the problem P can be solved as a linear complementarity problem, we need the following lemma which was used by Dantzig and Manne [2]. It is derived by taking limits of t-period sums.

LEMMA. *If* (x, y) *satisfies the following four conditions, then* x *solves* P.

1) $(A - B)x \leqq b, \quad x \geqq 0.$
2) $y(A - B)x = yb.$
3) $y(A - \delta B) \geqq v, \quad y \geqq 0.$
4) $y(A - \delta B)x = vx.$

We define the vector q and the matrix M as follows:

$$q = \begin{bmatrix} -v \\ b \end{bmatrix}, \qquad M = \begin{bmatrix} 0 & (A - \delta B)^t \\ B - A & 0 \end{bmatrix}.$$

A solution $z = (x, y)$ to the linear complementarity problem (q, M);

Find $z \in R^n$ such that $q + Mz \geqq 0$,
$z \geqq 0$,
$zq + zMz = 0$,
satisfies the four conditions of the lemma.

3. A termination result. In this section, we use the arguments of Dantzig and Manne [2] to establish a new termination criterion for Lemke's algorithm and show that it applies to the problem (q, M) defined in § 2. The result generalizes that of Dantzig and Manne [2]. We conclude by discussing a condition guaranteeing non-triviality of the computed solution.

Before stating and proving our main theorem, we briefly describe Lemke's algorithm. Lemke's algorithm is applied to the augmented system given in tableau form by

z	w	θ	
$-M$	I	$-e$	q

where e is a column vector of all ones. If $q \geqq 0$, then $w = q$, $\theta = 0$, $z = 0$ solves the problem. Otherwise, for θ sufficiently large, the system will be feasible with $z = 0$, $w = q + e\theta$. A starting basis is obtained by letting θ decrease until some component of w is driven to zero. This is done by performing a linear programming type pivot on row i of the θ column, where row i is chosen by picking the component q_i of q which is most negative. The basis then consists of θ and all w variables except w_i. The complement of w_i, z_i is then introduced to the basis using the usual min-ratio test and pivot of linear programming. After each pivot, either $\theta = 0$, in which case we have found a solution, or there is a distinguished pair (z_j, w_j), both terms of which are nonbasic and one of which has just left the basis. We then attempt to introduce the complement of the variable which just left the basis into the basis. If the min-ratio test fails to find an acceptable pivot row, then the new variable cannot be entered into the basis and we have terminated on a ray. As the algorithm cannot cycle, eventually either θ leaves the basis, in which case we have found a solution, or we encounter a ray.

THEOREM. *If the linear complementarity problem* (q, M) *satisfies*

1) $M + M^t \geqq 0$,

2) $q - M^t z \geqq 0$, $z \geqq 0$ *is feasible,*

then Lemke's algorithm applied to (q, M) *terminates in a solution.*

Proof. We apply the algorithm to the augmented system

$$w = q + Mz + e\theta,$$

where e is a column vector of ones. The algorithm stops if $\theta = 0$ (in which case we have obtained a solution) or if we encounter a ray. If we terminate in a ray, let (w_*, z_*, θ_*) denote the finite end of the ray and let $(w_h, z_h, \theta_h) \geqq 0$ denote the homogeneous part of the ray solution. That is, $w_h = Mz_h + e\theta_h$ and points along the ray are given parametrically by

$$w_r = (w_* + \lambda w_h), \qquad z_r = (z_* + \lambda z_h).$$

Note that by almost-complementarity we have

$$w_r z_r = (w_* + \lambda w_h)(z_* + \lambda z_h) = 0,$$

and we may then conclude that

$$w_* z_* = w_h z_* = w_* z_h = w_h z_h = 0. \tag{1}$$

From (1) we have that $0 = w_h z_h = z_h(Mz_h + e\theta_h)$. Since $z_h M z_h \geqq 0$, we find

$$z_h M z_h = \theta_h e z_h = 0. \tag{2}$$

From (2), we know that either $\theta_h = 0$ or $z_h = 0$. If both equal zero, then (w_h, z_h, θ_h) is a trivial homogeneous solution and could not be used to generate the ray. Suppose that $z_h = 0$ and $\theta_h \geqq 0$. Then $w_h \geqq 0$ and $w_h z_* = 0$ would imply that $z_* = 0$. The final ray would then be of the form $[z_r = 0, w_r = q + e\theta_r, \theta = \theta_r]$, where $\theta_r = \theta_* + \lambda\theta_h$. But this is the same ray as the initial ray which is a contradiction as Lemke's algorithm cannot cycle. It must then be the case that

$$\theta_h = 0. \tag{3}$$

From (3), we may then conclude that $z_* M z_h = z_*[Mz_h + e\theta_h] = z_* w_h$. From (1), we then have

$$z_* M z_h = 0. \tag{4}$$

Since $w_* = qMz_* + e\theta_*$, $z_h w_* = z_h q + z_h M z_* + z_h e\theta_*$, which equals 0 by (1). Equation (4) then allows us to write $0 = z_h q + z_h(M + M^t)z_* + z_h e\theta_*$, and since $M + M^t \geqq 0$ we may conclude that

$$z_h q \leqq 0, \qquad z_h M z_* \geqq 0. \tag{5}$$

Suppose now that $z_h q \leqq 0$. Since $Mz_h \geqq 0$ and $z_h \geqq 0$, we have by the Farkas lemma a violation of the second hypothesis of the theorem. Therefore,

$$z_h q = 0. \tag{6}$$

From equations (1), (5) and (6) and the fact that $w_* z_h = z_h q + z_h M z_* + z_h e\theta_*$, we obtain the conclusion that $\theta_* = 0$. Hence, we cannot terminate on a ray.

To apply the theorem to our problem, we need to guarantee that the two hypotheses are met. First note that $M + M^t \geqq 0$, because

$$M + M^t = (1 - \delta)\begin{bmatrix} 0 & B^t \\ B & 0 \end{bmatrix}$$

and $B \geqq 0$, $\delta \in (0, 1)$. The second hypothesis can be stated as:

There exists (x, y) such that

i) $y(A - B) \geqq v$, $y \geqq 0$,

ii) $(A - \delta B)x \leqq b$, $x \geqq 0$.

The duality theory of linear programming allows us to restate condition i) as

$$(A - B)x \leqq 0,\ x \geqq 0 \text{ implies } vx \leqq 0.$$

This is essentially a boundedness condition guaranteeing convergence of the infinite sums, and has the interpretation that any activity vector which is self-sufficient in that it uses no primary goods and produces all the production goods it uses must yield nonpositive utility to the economy.

The second condition is met by $x = 0$ if we restrict b to be nonnegative. Alternatively, condition ii) can be assumed directly. It has the interpretation that there exists a ray along which the economy could grow at a rate of $1/\delta$ if there were no restrictions upon the use of primary goods.

So far, we have not guaranteed that the computed solution is nontrivial, i.e., $vx > 0$. (Note that $x = 0$ may solve P.) To do this, we strengthen condition ii) and require the existence of $x \geqq 0$ such that $(A - \delta B)x < b$. This condition was first used by Peleg and Ryder [6] in a slightly different context and is called δ-productivity. It says that the economy could grow at a rate greater than $1/\delta$ if there were no restrictions upon the use of primary goods.

THEOREM. *If the technology is δ-productive and (x, y) satisfies the conditions of the lemma, then x is nontrivial.*

Proof. First note that since (x, y) satisfies the four conditions of the lemma, it must be the case that $vx \geqq yb$. Consider now the primal linear programming problem:

$$\min \quad yb \qquad \text{s.t.} \quad y(A-\delta B) \geqq v, \quad y \geqq 0.$$

By δ-productivity the dual problem has a feasible solution $\bar{x}$ such that $v\bar{x} > 0$. Applying the duality theorem of linear programming yields the conclusion that $yb > 0$.

4. The Leontief model. In this section, we examine a special case of the economic model which has a Leontief technology. Although the linear complementarity problem is in general NP-complete, the particular case examined here can be solved in $O(n^3)$ operations.

To fix ideas, we describe the Leontief consumption model. There are n production goods and 1 primary good, which we call labor. The technology has n productive activities, each of which produces exactly one production good and consumes various amounts of labor and production goods. We adopt the natural convention that the ith activity produces the ith good. There is an exogenously given supply of labor in each period which we take to equal 1. The utility is linear and is positive only for consumption activities. We have then

$$\bar{A} = \left[\begin{array}{c|c} A & I \\ \hline c & 0 \end{array}\right], \quad \bar{B} = \left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right], \quad \bar{b} = \left[\begin{array}{c} 0 \\ \hline 1 \end{array}\right], \quad \bar{v} = [0 \quad v],$$

where $c \in R^n_+$, $v \in R^n_+$ and $(I-A) \in R^{n \times n}$ is a Leontief matrix. That is, $(I-A)$ has nonpositive off-diagonal elements and positive principal minors. The paper by Fiedler and Pták [3] contains a discussion of such matrices.

The results of Fiedler and Ptak [3] show that δ-productivity implies that $(\delta I - A)$ is also a Leontief matrix. Thus $(I-A)$ and $(\delta I - A)$ have nonnegative inverses. A subscript i on a vector denotes its ith component. The solution is then found as follows. Let

$$\hat{c} = c(\delta 1 - A)^{-1}, \quad \tilde{c} = c(I-A)^{-1}, \quad \bar{w} = \max_i \left\{\frac{v_i}{\hat{c}_i}\right\};$$

pick

$$j \in \arg\max_i \left\{\frac{v_i}{\hat{c}_i}\right\},$$

and let $f = (f_1, f_2, \cdots, f_n)$ be given as

$$f_i = \begin{cases} 0 & \text{if } i \neq j, \\ \dfrac{1}{\tilde{c}_j} & \text{if } i = j. \end{cases}$$

Let $z = (I-A)^{-1} f$.

THEOREM. $x = (z, f)$ *as defined above solves the problem* P.

Proof. Let $y = (\bar{w}\hat{c}, \bar{w})$. The pair (x, y) then satisfies the sufficient conditions of the lemma.

At this point it is appropriate to note that f is obtained at the cost of solving two linear systems, n comparisons and $n+1$ divisions. The vector z can then be found using the factorization of $(I-A)$ already obtained.

Finally, we remark that for the Leontief consumption model examined in this section, it is not difficult to show that the conditions of the lemma are necessary as well as sufficient. Furthermore, if the model is irreducible, the only solutions are those given by the method shown above.

Acknowledgment. I would like to express my sincere appreciation of the advice afforded me by Professor David Gale who encouraged this work. Professor Ilan Adler's suggestions helped clarify several proofs. Additionally, I would like to thank Professor Richard Cottle who first aroused my interest in the linear complementarity problem. Finally, I would like to thank my colleague Pseb Vamajoth who proofread the manuscript and made several helpful suggestions.

REFERENCES

[1] S. J. CHUNG, *The linear complementarity problem is NP-complete*, Math. Programm., to appear.

[2] G. DANTZIG AND A. MANNE, *A complementarity algorithm for an optimal capital path with invariant proportions*, J. Econ. Theory, 9 (1974), pp. 312–323.

[3] M. FIEDER AND V. PTÁK, *On matrices with nonpositive off-diagonal elements and positive principal minors*, Czech. Math. J., 12 (1962), pp. 382–400.

[4] G. GALE, *On optimal development in a multi-sector economy*, Rev. Econ. Stud., 34 (1967), pp. 1–18.

[5] T. HANSEN AND T. KOOPMANS, *On the definition and computation of a capital stock invariant under optimization*, J. Econ. Theory, 5 (1972), pp. 487–523.

[6] B. PELEG AND H. RYDER, *The modified golden rule of a multi-sector economy*, J. Math. Econ., 1 (1974), pp. 193–198.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 2, June 1982

0196-5212/82/0302-0002 $01.00/0

SPREADS, TRANSLATION PLANES AND KERDOCK SETS. I*

WILLIAM M. KANTOR†

Abstract. In an orthogonal vector space of type $\Omega^+(4n, q)$, a spread is a family of $q^{2n-1}+1$ totally singular $2n$-spaces which induces a partition of the singular points; these spreads are closely related to Kerdock sets. In a $2m$-dimensional vector space over $GF(q)$, a spread is a family of q^m+1 subspaces of dimension m which induces a partition of the points of the underlying projective space; these spreads correspond to affine translation planes. By combining geometric, group theoretic and matrix methods, new types of spreads are constructed and old examples are studied. New Kerdock sets and new translation planes are obtained having various interesting properties.

1. Introduction. A Kerdock set is a collection of q^{2n-1} skew symmetric $2n \times 2n$ matrices over $GF(q)$ (with zero diagonal) the difference of any two of which is nonsingular. Each such set produces a spread of an $\Omega^+(4n, q)$ vector space; conversely, each spread produces at least one type of Kerdock set (up to a suitable version of equivalence). Certain Kerdock sets were first used by Kerdock [12] when $q = 2$ in order to construct nonlinear error-correcting codes having interesting properties (cf. Delsarte and Goethals [4], Goethals [9], MacWilliams and Sloane [14, Ch. 15]). Cameron and Seidel [1] give an elegant description of the relationship between Kerdock sets with $q = 2$ and codes.

Each translation plane over $GF(q)$ of order q^m can be coordinatized by an algebraic system which is equivalent to a set of q^m matrices, each of size $m \times m$, the difference of any two of which is nonsingular. Each such set produces a spread Σ of a $2m$-dimensional vector space V (Lüneburg [13, p. 8]). The plane can be recovered from V and Σ as follows: points are vectors, and lines are cosets $A + v$ with $A \in \Sigma$ and $v \in V$ [13, p. 2]. We will be concerned with a special type of spreads of V: symplectic spreads, in which there is an underlying $Sp\ (2m, q)$ structure on V such that each member of Σ is a totally isotropic m-space.

The similarity between the situations in the preceding paragraphs is obvious. For fields of characteristic 2, there is an elementary procedure for passing from orthogonal spreads to symplectic ones (Dillon [7], Dye [8]). This procedure is far from bijective: an $\Omega^+(4n, q)$ spread produces many different $Sp\ (4n-2, q)$ spreads, and hence many translation planes. Inequivalent orthogonal spreads never produce isomorphic translation planes.

The purpose of this paper is to construct new orthogonal spreads, and hence new Kerdock sets and translation planes; and to briefly study some of the translation planes arising from known orthogonal spreads. The main results can be summarized as follows.

The Kerdock sets originally constructed by Kerdock [12] and Delsarte and Goethals [4] correspond in a natural manner to desarguesian affine planes.

The $\Omega^+(2n, q)$ spreads (for q even) obtained from the desarguesian affine plane $AG(2, q^{2n-1})$ give rise to large numbers of pairwise nonisomorphic nondesarguesian translation planes of order q^{2n-1}, each admitting a collineation of order $q^{2n-1}+1$, transitively permuting the points at infinity. While the spreads defining these planes have been known implicitly for a while, the surprising fact that the planes are nondesarguesian was not.

* Received by the editors March 5, 1981.

† Bell Laboratories, Murray Hill, New Jersey 07974. Permanent address: Mathematics Department, University of Oregon, Eugene, Oregon 97403.

The unitary group $GU(3, q)$ (with $q \equiv 0$ or $2 \pmod 3$) gives rise to an $\Omega^+(8, q)$ spread Σ on which the projective unitary group $PGU(3, q)$ acts 2-transitively. When $q = 2^{2e+1} > 2$, one of the associated translation planes is a nondesarguesian plane of order q^3 admitting an $SL(2, q)$ acting with two orbits at infinity, of lengths $q+1$ and $q^3 - q$. The spread Σ also produces Kerdock sets of 4×4 skew symmetric matrices.

The group $SL(2, q^3)$ (with q even and $q > 2$) produces yet another $\Omega^+(8, q)$ spread, another Kerdock set, and two new translation planes.

Sections 2 and 3 contain background material. Section 3 contains the basic idea of the paper, and explains why our approach only produces translation planes of even order of the form $q^{2^{n-1}}$. That section also contains the crucial nonisomorphism criterion (3.6), which applies to all of the translation planes discussed throughout the paper. This criterion also makes it possible to determine the full collineation groups of our planes by using the more richly structured orthogonal spreads which produce the planes.

In § 4, "cousins" of desarguesian planes are constructed. Section 5 describes the elementary procedure for passing between orthogonal spreads and Kerdock sets.

Section 6 is concerned with a representation of $GU(3, q)$ (essentially the adjoint representation) and the corresponding spread (when $q \equiv 0$ or $2 \pmod 3$); some of the resulting translation planes are discussed in § 7. An irreducible 8-dimensional representation of $SL(2, q^3)$ produces spreads and planes in § 8, while an isolated $GF(8)$ example due to Dye [8] is discussed in § 9.

Section 10 summarizes all the previous results in terms of Kerdock sets. Automorphism groups are listed; they are subgroups of the groups of the corresponding spreads. The ease of dealing with automorphism groups is an example of the advantage of "extending" a Kerdock set to a spread: the latter has a richer structure. (This is reminiscent of adding an overall parity check to a code.) Perfect 1-codes in certain graphs are discussed in § 11.

In a sequel we will construct spreads in orthogonal spaces of arbitrarily large dimension over any field of characteristic 2. Unfortunately, despite remarks in [19, § 2(b)] and [20, § 1(b)], examples are not even known in $\Omega^+(8, q)$ spaces when $q \equiv 1 \pmod 6$.

2. Preliminaries. We begin with a brief review of orthogonal geometry (cf. Dieudonné [6]). Let V be a $2m$-dimensional vector space over a field K. A quadratic form Q on V, with associated symmetric bilinear form $(\cdot, \cdot)$, satisfies $Q(av) = a^2 Q(v)$ and $Q(u+v) - Q(u) - Q(v) = (u, v)$ for all $u, v \in V$ and $a \in K$. The form (u, v) will be nonsingular, unless stated otherwise (cf. (6.3), (6.8)).

A singular vector v satisfies $Q(v) = 0$, while a totally singular subspace W satisfies $Q(W) = 0$. Here, $\dim W \leqq m$. We will be concerned with spaces of type $\Omega^+(2m, K)$, in which totally singular m-spaces exist. For such a space, there are two types of totally singular m-spaces; two have the same type if and only if the dimension of their intersection has the same parity as m [6, pp. 50, 65, 86, 87]. Each totally singular $(m-1)$-space is contained in exactly two totally singular m-spaces, one of each type.

If V is an $\Omega^+(2, K)$ space, it is called a hyperbolic line. If $2m = 2$ and V is not a hyperbolic line, it is called anisotropic; there are then no nonzero singular vectors.

An orthogonal $2m$-space is an $\Omega^+(2m, K)$ space if and only if it is the direct sum of pairwise orthogonal hyperbolic lines.

A point is, of course, just a 1-space. It may be singular or nonsingular.

If $\operatorname{char} K = 2$, then V is symplectic as well as orthogonal: $y \leqq y^\perp$ for each point y. Then $y^\perp / y$ is another nonsingular symplectic space. In particular, if y is nonsingular then the natural map $y^\perp \to y^\perp / y$ from the $(2m-1)$-dimensional orthogonal space $y^\perp$ to

the $(2m-2)$-dimensional symplectic space $y^\perp/y$ bijectively maps totally singular subspaces to totally isotropic subspaces, and $\Omega(2m-1, K) \cong Sp(2m-2, K)$ if K is perfect.

$\Gamma L(2m, K)$ is the group of invertible semilinear transformations of V, while $\Gamma O^+(2m, K)$ is its subgroup preserving Q projectively. (Thus, each $g \in \Gamma O^+(2m, K)$ determines a $c \in K$ and a $\phi \in \mathrm{Aut}\, K$ such that $Q(v^g) = cQ(v)^\phi$ for all $v \in V$.) The group $\Gamma Sp(2m-2, K)$ is defined similarly.

If $G \leqq \Gamma L(2m, K)$ and Σ is a family of subspaces of V, then

$$G_\Sigma = \{g \in G | \Sigma^g = \Sigma\}.$$

Also, $C_V(G) = \{v \in V | v^g = v \text{ for all } g \in G\}$.

3. Spreads and translation planes. Let V be an $\Omega^+(4n, q)$ vector space with q even. Let Σ be a spread of V. Then Σ consists of $q^{2n-1}+1$ totally singular $2n$-spaces, and each nonzero singular vector lies in exactly one of them. If y is any nonsingular point, then

$$(y^\perp \cap \Sigma)/y = \{\langle y, y^\perp \cap F\rangle/y | F \in \Sigma\} \tag{3.1}$$

is a spread of the symplectic space $y^\perp/y$. (For, each nonzero singular vector of $y^\perp$ belongs to exactly one space $y^\perp \cap F$; moreover, $y^\perp$ cannot contain any F so that $\dim y^\perp \cap F = 2n-1$.)

Conversely, let Σ' be a spread of a symplectic space V' of dimension $4n-2$ over $GF(q)$, where q is even. Identify V' with $y^\perp/y$, for a nonsingular point y of an $\Omega^+(4n, q)$ space V. Fix a type of totally singular $2n$-space of V. Set

(3.2)

$$\mathscr{S}(\Sigma') = \{F | F \text{ is a totally singular } 2n\text{-space of the fixed type, and } \langle y, y^\perp \cap F\rangle/y \in \Sigma'\}.$$

Then $\Sigma = \mathscr{S}(\Sigma')$ is a spread of V. (For, each $F' \in \Sigma'$ arises as $\langle y, y^\perp \cap F\rangle/y$ for a unique F of the fixed type. If $F_1, F_2 \in \mathscr{S}(\Sigma')$, $F_1 \neq F_2$, then $(y^\perp \cap F_1) \cap (y^\perp \cap F_2) = 0$ since Σ' is spread. However, $\dim (F_1 \cap F_2) \equiv 2n \pmod 2$. Thus, $F_1 \cap F_2 = 0$, and hence $|\cup\{F - \{0\} | F \in \Sigma\}| = (q^{2n-1}+1)(q^{2n}-1)$, so Σ is indeed a spread of V.)

Clearly,

$$\Sigma' = (y^\perp \cap \mathscr{S}(\Sigma'))/y \quad \text{and} \quad \Sigma = \mathscr{S}((y^\perp \cap \Sigma)/y). \tag{3.3}$$

The preceding constructions are due to Dillon [7] and Dye [8].

With each spread Σ' of V' is associated a translation plane $\mathscr{A}(\Sigma')$ of order q^{2n-1}, which was defined in § 1. Here,

$$\mathrm{Aut}\, \mathscr{A}(\Sigma') = V' \rtimes \Gamma L(4n-2, q)_{\Sigma'}. \tag{3.4}$$

Note that $\Gamma L(4n-2, q)_{\Sigma'}$ can be larger than $\Gamma Sp(4n-2, q)_{\Sigma'}$ (cf. § 4).

If Σ' and Σ'' are spreads of the symplectic space V', and are equivalent under the action of $\Gamma Sp(4n-2, q)$, then $\mathscr{S}(\Sigma')$ and $\mathscr{S}(\Sigma'')$ are equivalent under the action of $\Gamma O^+(4n, q)$ (but not conversely). Moreover, $\mathscr{A}(\Sigma')$ and $\mathscr{A}(\Sigma'')$ are isomorphic; remarkably, the converse is true.

THEOREM 3.5. *Let Σ_i be a spread of an $Sp(2m, q)$ space V_i $(i = 1, 2)$, where q is even. If $\mathscr{A}(\Sigma_1) \cong \mathscr{A}(\Sigma_2)$, then there is a semilinear transformation $s: V_1 \to V_2$ such that*

(i) $\Sigma_1^s = \Sigma_2$, *and*

(ii) $(u^s, v^s)_2 = a(u, v)_1^\tau$ *for some $a \in GF(q)$, some $\tau \in \mathrm{Aut}\, GF(q)$, and all $u, v \in V_1$.*

Proof. Let θ be the polarity of $PG(2m-1, q)$ determined by the symplectic structure on V_1 (so that $W^\theta = W^\perp$ for each subspace W of V_1). Define ϕ similarly for V_2. We are given a semilinear transformation $g: V_1 \to V_2$ such that $\Sigma_1^g = \Sigma_2$. Set $\theta^g = g^{-1}\theta g$.

If $F \in \Sigma_1$ then $F^\theta = F$ and $(F^g)^\phi = F^g$. Moreover, $(F^g)^{\theta^g} = F^{gg^{-1}\theta g} = F^{\theta g} = F^g$. Thus, ϕ and θ^g are polarities of the projective space corresponding to V_2, both of which induce the identity on Σ_2. Then $\phi\theta^g$ is a collineation of that projective space, and hence is induced by a semilinear transformation h of V_2. Here, h is the identity on Σ_2, and hence is an homology of $\mathscr{A}(\Sigma_2)$. In particular, $|h|$ is odd since q is even (Dembowski [5, p. 172]).

Set $|\phi\theta^g| = 2j+1$ and $f = (\phi\theta^g)^j$. Then $(\theta^g)^f = \phi$ since $\phi^2 = 1$ and $\theta^2 = 1$. Consequently, $\Sigma_1^{gf} = \Sigma_2^f = \Sigma_2$ and $\theta^{gf} = \phi$. There is thus a semilinear transformation $s : V_1 \to V_2$ inducing gf and satisfying (ii) (Dieudonné [6, Ch. III, § 3]); clearly, s also satisfies (i). □

Remarks. The above proof provides some insight into the nature of a symplectic spread: the polarity θ is trying to act on $\mathscr{A}(\Sigma_1)$ as an homology.

If $\Sigma_1 = \Sigma_2$, then (ii) asserts that $s \in \Gamma Sp(2m, q)_{\Sigma_1}$.

COROLLARY 3.6. *Let Σ_i be a spread in an $\Omega^+(4n, q)$ space V_i $(i = 1, 2)$, where q is even. Let y_i be a nonsingular point of V_i. Assume that $\mathscr{A}((y_1^\perp \cap \Sigma_1)/y_1) \cong \mathscr{A}((y_2^\perp \cap \Sigma_2)/y_2)$. Then there is a semilinear transformation $h : V_1 \to V_2$ such that*

(i) $\Sigma_1^h = \Sigma_2$,

(ii) $y_1^h = y_2$, *and*

(iii) $Q_2(v^h) = aQ_1(v)^\tau$ *for some* $a \in GF(q)$, *some* $\tau \in \mathrm{Aut}\, GF(q)$, *and all* $v \in V_1$.

Proof. By Theorem 3.5, there is a semilinear transformation $h : y_1^\perp \to y_2^\perp$ such that $(y_1^\perp \cap \Sigma_1)^h = y_2^\perp \cap \Sigma_2$, $y_1^h = y_2$, and (iii) holds for some a, some τ and all $v \in y_1^\perp$. There are exactly two extensions of h to a map $V_1 \to V_2$ satisfying (ii) and (iii), only one of which maps the type of $2n$-space in Σ_1 to the type of $2n$-space in Σ_2. In view of (3.3), the latter extension satisfies (i)–(iii). □

COROLLARY 3.7. *If Σ, Σ' and y are as in* (3.3) *then every collineation of $\mathscr{A}(\Sigma')$ fixing* 0 *can be written in the form sf, where s is induced by an element of $\Gamma O^+(4n, q)_{\Sigma, y}$ and f is an homology of $\mathscr{A}(\Sigma')$.*

Proof. This is implicit in the proof of Theorem 3.5. □

4. Desarguesian spreads and their cousins. If $[\cdot, \cdot]$ is a nonsingular symplectic form on a 2-dimensional $GF(q^m)$-space V, then $(u, v) = T[u, v]$ defines a nonsingular symplectic form on the $GF(q)$-space V for any nonzero $GF(q)$-linear functional T on $GF(q^m)$. One-dimensional $GF(q^m)$-subspaces become totally isotropic m-spaces, and we obtain a spread of V. Any symplectic spread obtained in this manner will be called *desarguesian*: the corresponding translation plane is just the desarguesian one $AG(2, q^m)$ (Lüneburg [13]).

If q is even and $m = 2n-1$, then we obtain a spread Σ' of an $Sp(4n-2, q)$ space, and hence can apply the previous section. The spread $\mathscr{S}(\Sigma')$ will be called a *desarguesian spread of an $\Omega^+(4n, q)$ space.*

DEFINITION. Two spreads Σ' and Σ'' of an $Sp(4n-2, q)$ space (with q even) are called *cousins* if $\mathscr{S}(\Sigma')$ and $\mathscr{S}(\Sigma'')$ are equivalent under the action of $\Gamma O^+(4n, q)$. In this case the translation planes $\mathscr{A}(\Sigma')$ and $\mathscr{A}(\Sigma'')$ will also be called cousins.

It should be noted that the process of passing from a translation plane to one of its cousins is quite different from that of derivation or net replacement (Lüneburg [13]). In fact, the planes we are considering have order q^{2n-1}, which need not even be a square.

Given a spread Σ' of an $Sp(4n-2, q)$ space V' (with q even), all cousins are "found" as follows. Form the spread $\Sigma = \mathscr{S}(\Sigma')$ as in (3.2). Then construct the various slices (3.1) of Σ. Of course, all that is really desired are slices that are inequivalent under the action of $\Gamma O^+(4n, q)$. This naturally leads to the consideration of the orbits of nonsingular points under the action of $\Gamma O^+(4, q)_\Sigma$. When Σ' is desarguesian, the latter

group was determined by Dye [8] and Cohen and Wilbrink [2]. It is then straightforward to determine the orbits of $\Gamma O^+(4n, q)_\Sigma$ on nonsingular points. This was done by Dye [8] in the case of $O^+(4n, q)_\Sigma$ orbits. For completeness, we will describe these orbits.

LEMMA 4.1. *Let Σ' be a desarguesian spread of an $Sp(4n-2, q)$-space, where q is even. Define V, y and $\Sigma = \mathscr{P}(\Sigma')$ as in (3.2). Exclude the case $n = q = 2$. Then $G = \Gamma O^+(4n, q)_\Sigma \cong \Gamma\ Sp\ (4n-2, q)_{\Sigma'} = (SL(2, q^{2n-1}) \times GF(q)^*) \cdot \mathrm{Aut}\ (GF(q^{2n-1}))$. The cousins of Σ' arise as $\Sigma'' = (y''^{\perp} \cap \Sigma)/y''$, with y'' one of the following types of nonsingular points (where $l = \log_2 q$).*

(I) $y'' = y$.

(II) $y \neq y'' < y^{\perp}$, $|G_y| = q^{2n-1}(q-1)(2n-1)l$ *(one G-orbit); G_y fixes one member of Σ (which meets $\langle y, y''\rangle$ nontrivially) and is transitive on the remaining ones.*

(III) *$\langle y, y''\rangle$ is a hyperbolic line, $|G_{y''}| = (q^{2n-1}-1)(q-1)(2n-1)l''$ with $l''|l$ (at least $(\frac{1}{2}q-1)/l$ such orbits); $G_{y''}$ has three orbits on Σ, of lengths 1, 1 and $q^{2n-1}-1$.*

(IV) *$\langle y, y''\rangle$ is an anisotropic line, $|G_{y''}| = (q^{2n-1}-1)(q-1)(2n-1)l''$ with $l''|l$ (at least $\frac{1}{2}q/l$ such orbits); $G_{y''}$ has a cyclic subgroup transitive on Σ.*

Proof. That $G \cong \Gamma Sp\ (4n-2,\ q)_{\Sigma'} = (SL(2, q^{2n-1}) \times GF(q)^*) \cdot \mathrm{Aut}\ GF(q^{2n-1})$ is essentially contained in [8] and [2]. Thus, G fixes y; consider its action on $y^{\perp}$. It is transitive on the hyperplanes of $y^{\perp}$ through y, on the singular points of $y^{\perp}$, and on the nonsingular points other than y. There are thus 3 point-orbits on $y^{\perp}$, and hence also 3 hyperplane-orbits (Dembowski [5, p. 78]), which can only be the hyperplanes through y and the nonsingular hyperplanes H of $y^{\perp}$ of each type. Note that $H^{\perp}$ is either a hyperbolic line or an anisotropic line through y. This accounts for all the types (I)–(IV). Keeping in mind the fact that G contains all scalar transformations, the lemma follows easily. □

THEOREM 4.2. *Let q be even and $q^n > 8$, where $n > 1$. Let Σ be a desarguesian spread of an $\Omega^+(4n, q)$ space, and consider the resulting cousins of $AG(2, q^{2n-1})$.*

(i) *If two cousins $\mathscr{A}((y_i^{\perp} \cap \Sigma)/y_i)$ are isomorphic $(i = 1, 2)$, then y_1 and y_2 are $\Gamma O^+(4n, q)_\Sigma$-equivalent.*

(ii) *The cousin arising from Lemma 4.1 (II) is a nondesarguesian semifield plane of order q^{2n-1}.*

(iii) *The cousins arising from Lemma 4.1 (III) are nondesarguesian of order q^{2n-1}. There are at least $(q-2)/(2\log_2 q)$ of these cousins. The full collineation group of each induces a subgroup of $GF(q^{2n-1})^* \cdot \mathrm{Aut}\ GF(q^{2n-1})$ on the line at infinity, with orbit lengths 1, 1, $q^{2n-1}-1$.*

(iv) *The cousins arising from Lemma 4.1 (IV) are flag-transitive nondesarguesian planes of order q^{2n-1}. There are at least $q/(2\log_2 q)$ pairwise nonisomorphic cousins of this type. The full collineation group of each of them induces a group of order dividing $(q^{2n-1}+1)\log_2 q^{2n-1}$ on the line at infinity, having a normal cyclic subgroup transitive on the line at infinity.*

Proof. By Corollary 3.6, (i) holds. All remaining assertions are immediate consequences of Lemma 4.1 and Corollary 3.7. □

Further information concerning these cousins will not be required in this paper, and hence is postponed until a subsequent paper.

5. Kerdock sets. Fix an $\Omega^+(4n, q)$ space V (where q may be even or odd), and two totally singular $2n$-spaces E and F such that $V = E \oplus F$. There are bases $e_1, \cdots, e_{2n}$ and $f_1, \cdots, f_{2n}$ of E and F such that $(e_i, f_j) = \delta_{ij}$ for all i, j. Of course, $Q(e_i) = Q(f_j) = 0$.

Writing matrices with respect to the ordered basis $e_1, \cdots, e_{2n}, f_1, \cdots, f_{2n}$, we find that

$$\begin{pmatrix} I & O \\ M & I \end{pmatrix} \text{ preserves } Q \Leftrightarrow M^t = -M \text{ and } M \text{ has zero diagonal,} \tag{5.1}$$

where O and I denote the $2n \times 2n$ zero and identity matrices. Thus, the group $P^\natural$ of matrices (5.1) preserving Q is isomorphic to the group P of skew symmetric matrices (with zero diagonal), via $M \to M^\natural = \left(\begin{smallmatrix} I & O \\ M & I \end{smallmatrix}\right)$ for $M \in P$.

A Kerdock set is a subset $\mathcal{K}$ of P consisting of q^{2n-1} matrices the difference of any two of which is nonsingular. Clearly, $\mathcal{K}$ determines a subset $\mathcal{K}^\natural$ of P. Define

$$\mathcal{S}(\mathcal{K}) = \{E\} \cup \{F^g \mid g \in \mathcal{K}^\natural\}. \tag{5.2}$$

If $g = \left(\begin{smallmatrix} I & O \\ M & I \end{smallmatrix}\right)$ and $g' = \left(\begin{smallmatrix} I & O \\ M' & I \end{smallmatrix}\right)$ belong to $P^\natural$, then $F^g \cap F^{g'} = 0$ precisely when $M - M'$ is nonsingular. Thus, $\mathcal{S}(\mathcal{K})$ *is a spread of* V.

Conversely, let Σ be any spread of V. Pick *any* two members of Σ, call them E and F, and pick a basis $e_1, \cdots, e_{2n}, f_1, \cdots, f_{2n}$ as above. Define

$$K_E(\Sigma) = \{M \in P \mid FM^\natural \in \Sigma - \{E\}\}. \tag{5.3}$$

Then $|K_E(\Sigma)| = |\Sigma| - 1 = q^{2n-1}$, and hence $K_E(\Sigma)$ *is a Kerdock set* since any two members of Σ intersect trivially.

Clearly, $K_E(\mathcal{S}(\mathcal{K})) = \mathcal{K}$ and $\Sigma = \mathcal{S}(K_E(\Sigma))$, assuming that we fix our basis $e_1, \cdots, e_{2n}, f_1, \cdots, f_{2n}$.

Two Kerdock sets $\mathcal{K}$ and $\mathcal{K}_1$ in P are called *equivalent* if there is a transformation $M \to dA^{-1}M^\phi(A^{-1})^t + C$ sending $\mathcal{K}$ to $\mathcal{K}_1$, where $d \in GF(q)^*$, $A \in GL(2n, q)$, $\phi \in \mathrm{Aut}\, GF(q)$ and $C \in P$.

LEMMA 5.4. *Let $\mathcal{K}$ and $\mathcal{K}_1$ be Kerdock sets in P. Then $\mathcal{K}$ and $\mathcal{K}_1$ are equivalent if and only if there is an element $g \in \Gamma O^+(4n, q)_E$ such that $\mathcal{S}(\mathcal{K})^g = \mathcal{S}(\mathcal{K}_1)$.*

Proof. Let $g \in \Gamma O^+(4n, q)_E$. Then $v^g = v^\phi \left(\begin{smallmatrix} A & O \\ O & B \end{smallmatrix}\right)\left(\begin{smallmatrix} I & O \\ C & I \end{smallmatrix}\right)$ for $2n \times 2n$ matrices A, B, C, and $\phi \in \mathrm{Aut}\,(GF(q))$, while $Q(v^g) = dQ(v)^\psi$ for some $d \in GF(q)^*$ and $\psi \in \mathrm{Aut}\, GF(q)$. Then $\psi = \phi$, while $AB^t = dI$ and $C \in P$.

If $F\left(\begin{smallmatrix} I & O \\ N & I \end{smallmatrix}\right) \in \mathcal{S}(\mathcal{K})$, so $N^\natural \in \mathcal{K}$, then

$$\left(F\begin{pmatrix} I & O \\ N & I \end{pmatrix}\right)^g = F\begin{pmatrix} I & O \\ N^\phi & I \end{pmatrix}\begin{pmatrix} A & O \\ O & B \end{pmatrix}\begin{pmatrix} I & O \\ C & I \end{pmatrix}$$

$$= F\begin{pmatrix} A^{-1} & O \\ O & B^{-1} \end{pmatrix}\begin{pmatrix} I & O \\ N^\phi & I \end{pmatrix}\begin{pmatrix} A & O \\ O & B \end{pmatrix}\begin{pmatrix} I & O \\ C & I \end{pmatrix} = F\begin{pmatrix} I & O \\ A^{-1}N^\phi B + C & I \end{pmatrix}.$$

Thus $\mathcal{S}(\mathcal{K})^g = \mathcal{S}(\mathcal{K}_1)$ if and only if $N \to dA^{-1}N^\phi(A^{-1})^t + C$ takes $\mathcal{K}$ to $\mathcal{K}_1$. □

Note that (5.4) shows that $K_E(\Sigma)$ depends on neither F nor the choice of basis $e_1, \cdots, e_{2n}$ of E (up to equivalence). Also, by Lemma 5.4 the determination of all Kerdock sets in P is equivalent to the determination of all pairs (Σ, E) consisting of a spread Σ of V and a member E of Σ. Examples Σ exist for which $G = \Gamma O^+(4n, q)_\Sigma$ is not transitive on Σ, and hence which produce at least two inequivalent Kerdock sets.

However, if Σ' is a desarguesian spread of an $Sp(4n-2, q)$-space with q even (§ 4), then $\Sigma = \mathcal{S}(\Sigma')$ has G acting 3-transitively on Σ as $P\Gamma L(2, q^{2n-1})$. The corresponding Kerdock set $K_E(\mathcal{S}(\Sigma'))$ is then independent of E. This is precisely the usual Kerdock set, discovered by Kerdock [12] and Delsarte and Goethals [4] (compare MacWilliams and Sloane [14, Ch. 15]). It will be called the *desarguesian Kerdock set.*

6. Unitary spreads. In this section, spreads will be constructed using the unitary groups $GU(3, q)$. It will be convenient to avoid finiteness initially and to start with any field F of characteristic p.

The group $G_0 = GL(3, F)$ acts by conjugation on the space $V_0 = sl(3, F)$ of 3×3 matrices of trace 0. If $M, N \in V_0$ then $(M, N) = \mathrm{tr}\,(MN)$ defines a symmetric bilinear

form preserved by G. If $M = (\mu_{ij})$, define

$$Q(M) = -\sum_{i<j} \mu_{ii}\mu_{jj} + \sum_{i<j} \mu_{ij}\mu_{ji}. \tag{6.1}$$

Then $Q(M+N) - Q(M) - Q(N) = (M, N)$, so that Q is a quadratic form with associated bilinear form $(\cdot\,, \cdot)$.

LEMMA 6.2. *G_0 preserves Q.*

Proof. If $p \neq 2$ then $(M, M) = 2Q(M)$, and the lemma is obvious.

The case $p = 2$ can be handled by calculation, or as follows. Let R be a commutative ring in which 2 is invertible. Adjoin indeterminates μ_{ij}, a_{ij} and b_{ij} $(1 \leqq i, j \leqq 3)$ subject to the relations $\sum_i \mu_{ii} = 0$ and $AB = I$, where $A = (a_{ij})$ and $B = (b_{ij})$. Set $M = (\mu_{ij})$. If Q is defined by (6.1), then $Q(BMA) = Q(M)$ once again. Now pass mod 2 and specialize the indeterminates in order to deduce the lemma. □

LEMMA 6.3. *If $p \neq 3$ then* rad $V_0 = 0$. *If $p = 3$ then* rad $V_0 = \langle I \rangle$, *where I is the identity matrix.*

Proof. If E_{ij} is the matrix having (i, j)-entry 1 and all other entries 0, then tr (AE_{ij}) is the (i, j)-entry of A. Let $A = (a_{ij}) \in$ rad V. Then $a_{ij} = 0$ for $i \neq j$. Also, $0 =$ tr $(A(E_{ii} - E_{jj}))$. Thus, $A = a_{11}I$, as required. □

LEMMA 6.4. *Let $p \neq 3$. Then V_0 has type $\Omega^+(8, F)$ if and only if F contains a primitive cube root of unity.*

Proof. The matrices

$$\begin{pmatrix} 0 & a & 0 \\ b & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & c \\ 0 & 0 & 0 \\ d & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & e \\ 0 & f & 0 \end{pmatrix}, \quad \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \gamma \end{pmatrix}$$

are pairwise orthogonal, and produce a decomposition of V_0 into four nonsingular 2-spaces. On the first three of these, Q induces the form xy, while on the last it induces the form $-\alpha\beta - \alpha(-\alpha - \beta) - \beta(-\alpha - \beta) = \alpha^2 + \alpha\beta + \beta^2$. Thus, V_0 has type $\Omega^+(8, F)$ if and only if $x^2 + x + 1 = 0$ has a root. □

We now turn to transvections (i.e., elations of $PSL(3, F)$). A nontrivial transvection is defined to be a matrix of the form $I + X$ with $X^2 = 0 \neq X$. Clearly, such a matrix X belongs to V_0. A full transvection group consists of all transvections with a given center and axis. (In terms of the desarguesian plane $PG(2, F)$, the axis is the line fixed pointwise and the center is the point fixed linewise; cf. Dembowski [5, p. 119].) Such a group has the form $\{I + aX \mid a \in F\}$. All nontrivial transvections are conjugate [6], so $Q(X) = Q(E_{13}) = 0$.

Let Φ denote the set of all the corresponding singular points $\langle X \rangle$.

LEMMA 6.5. *If $\langle X \rangle \in \Phi$, set $T_0(X) = \{M \in V_0 \mid XM = MX = 0\}$. Then* dim $T_0(X) =$ 3 *and $Q(T_0(X)) = 0$. (Moreover, if $p \neq 3$ then $T_0(X)$ consists of all singular vectors in $C_{V_0}(X)$.)*

Proof. If $X = E_{13}$ then $C_{V_0}(X)$ consists of all matrices

$$M = \begin{pmatrix} a & b & c \\ 0 & -2a & d \\ 0 & 0 & a \end{pmatrix},$$

with $a, b, c, d \in F$. Here, $MX = 0$ if and only if $a = 0$, while $Q(M) = -a(-2a) - a \cdot a - (-2a)a = 3a^2$. Thus,

$$T_0(X)=\left\{\begin{pmatrix}0&b&c\\0&0&d\\0&0&0\end{pmatrix}\middle|\, b,c,d\in F\right\}\quad\text{if } X=\begin{pmatrix}0&0&1\\0&0&0\\0&0&0\end{pmatrix}. \tag{6.6}$$

This proves the lemma. □

Remark. $T_0(aX)=T_0(X)$ if $a\neq 0$.

LEMMA 6.7. *If* $\langle X\rangle,\langle Y\rangle\in\Phi$ *and* $T_0(X)\cap T_0(Y)\neq 0$, *then* $T_0(X)\cap T_0(Y)$ *contains some* $\langle Z\rangle\in\Phi$.

Proof. $\langle X\rangle$ corresponds to a flag in $PG(2,F)$ (namely, the center-axis flag of $I+X$). By (6.6), $T_0(X)$ "contains" every flag having the same center or axis as $I+X$. If the center of $I+X$ lies on the axis of $I+Y$, or vice versa, then $T_0(X)$ and $T_0(Y)$ "contain" a common flag. We can now use the transitivity of $(G_0)_{\langle X\rangle}$ in order to specialize Y to $Y=X^t$; but then $T_0(Y)=T_0(X)^t$ by definition, while $T_0(X)\cap T_0(X)^t=0$ by (6.6). □

From now on, we specialize to the case $F=GF(q^2)$. Set $\alpha^q=\bar\alpha$ for $\alpha\in F$, and write $K=GF(q)$. Let $G=GU(3,q)$ be the unitary group consisting of all matrices $A\in GL(3,q^2)$ such that $A^{-1}=\bar A^t$, and set

$$V=\{M\in V_0\,|\,\bar M^t=M\}.$$

Then $\dim_K V=8$.

Note that G acts on V (since $A^{-1}MA=(\bar A^{-1}\bar M\bar A)^t$).

If $M=(\mu_{ij})\in V$, then $Q(M)$ is still given by (6.1). Since $\mu_{ii}\in K$ and $\mu_{ji}=\bar\mu_{ij}$, we have $Q(M)\in K$. Thus, Q is a quadratic form $V\to K$. Its associated bilinear form is just the restriction of $(\cdot\,,\cdot)$ to $V\times V$.

LEMMA 6.8. *If* $p\neq 3$ *then* rad $V=0$. *If* $p=3$ *then* rad $V=\langle I\rangle$.

Proof. Compute! □

LEMMA 6.9. *Let* $p\neq 3$. *Then* V *has type* $\Omega^+(8,q)$ *if and only if* $q\equiv 2\pmod 3$.

Proof. Set

$$[a,b;\alpha]=\begin{pmatrix}a&\bar\alpha&0\\\alpha&b&0\\0&0&-a-b\end{pmatrix}\quad\text{and}\quad[\beta,\gamma]=\begin{pmatrix}0&0&\bar\beta\\0&0&\bar\gamma\\\beta&\gamma&0\end{pmatrix} \tag{6.10}$$

$$V_1=\{[a,b;\alpha]\,|\,a,b\in K,\alpha\in F\}\quad\text{and}\quad V_2=\{[\beta,\gamma]\,|\,\beta,\gamma\in F\}. \tag{6.11}$$

Then $V=V_1\perp V_2$. The matrices $[a,b;0]$, $[0,0;\alpha]$, $[\beta,0]$ and $[0,\gamma]$ are pairwise orthogonal, and hence produce nonsingular 2-spaces. The last three 2-spaces are anisotropic lines. The first is anisotropic if and only if $Q([a,b;0])=a^2+ab+b^2$ is never 0 for $ab\neq 0$, which is true precisely when $GF(q)$ does not contain a primitive cube root of unity, i.e., when $q\equiv 2\pmod 3$. This proves the lemma. □

The transvections in G have the form $I+Y$ with $Y^2=0$ and $I-Y=(I+Y)^{-1}=\bar I^t+\bar Y^t$. Fix $\theta\in F$ with $\bar\theta=-1$. (If $p=2$, set $\theta=1$. If $p\neq 2$ then θ can be taken as $(\phi-\bar\phi)/2$ for any $\phi\in F-K$.) Then $I+Y=I+\theta X$ with $X^2=0\neq X$ and $X\in V$.

Set $\Omega=\{\langle X\rangle\,|\,X^2=0\neq X, X\in V\}$. Then Ω consists of singular points of V which are permuted 2-transitively by G (Dembowski [5, p. 54]; Lüneburg [13, p. 157]).

LEMMA 6.12. *If* $\langle X\rangle\in\Omega$, *let* $T(X)=\{M\in V\,|\,XM=MX=0\}$. *Then*

(i) $\dim T(X)=3$,

(ii) $T(X)$ *consists of singular vectors, and*

(iii) $T(X)\cap T(Y)=0$ if $\langle X\rangle\neq\langle Y\rangle\in\Omega$.

Proof. (i) Clearly, $T(X) = V \cap T_0(X)$. If $\alpha \in F - K$ then $T_0(X) \supseteq T(X) + \alpha T(X)$, so $6 = \dim_K T_0(X) \geqq 2 \dim_K T(X)$.

Let $p = 2$. Then $M \to \bar{M}^t$ is an involutory K-linear transformation on $T_0(X)$. Hence, its set of fixed points has dimension $\geqq \frac{1}{2} \dim_K T_0(X)$, proving (i) in this case.

If $p \neq 2$, recall that $\bar{\theta} = -\theta$. If $M \in T_0(X)$ then $M = \frac{1}{2}(M + \bar{M}^t) + \frac{1}{2}(M - \bar{M}^t)$, while $\theta T(X) = \{N \in T_0(X) | \bar{N}^t = -N\}$. Thus, $\dim_K T(X) = \dim_K \theta T(X) = \frac{1}{2} \dim_K T_0(X) = 3$.

(ii) See Lemma 6.5.

(iii) Since no two members of Ω can be perpendicular, this follows from (6.6) (compare Dembowski [5, p. 104]; Lüneburg [13, p. 154]). □

DEFINITION. If $q \equiv 2 \pmod 3$, fix a type of totally singular 4-space of V (cf. Lemma 6.9). Let $F(X)$ be the subspace of that type containing $T(X)$. Set $\Sigma = \{F(X) | \langle X \rangle \in \Omega\}$.

THEOREM 6.13. (i). *If* $q \equiv 2 \pmod 3$, *then* Σ *is a spread of the* $\Omega^+(8, q)$ *space* V *permuted* 2*-transitively by* $PGU(3, q)$.

(ii) *If* $q \equiv 0 \pmod 3$ *then* $\{\langle T(X), I \rangle / \langle I \rangle | \langle X \rangle \in \Omega\}$ *is a spread of the* $\Omega(7, q)$ *space* $V/\langle I \rangle$ *permuted* 2*-transitively by* $PGU(3, q)$.

Proof. (i) Two totally singular subspaces of the chosen type have intersection of dimension 0, 2 or 4. By Lemma 6.12 (iii), distinct members of Σ intersect trivially. They thus cover $|\Sigma|(q^4 - 1)/(q - 1) = (q^3 + 1)(q^2 + 1)(q + 1)$ singular points, that is, all the singular points of V.

(ii) This time $V/\langle I \rangle$ has exactly $|\Omega|(q^3 - 1)/(q - 1)$ singular points. □

DEFINITION. Let $q \equiv 0 \pmod 3$. Let $V^{\#}$ be an $\Omega^+(8, q)$ space containing $V/\langle I \rangle$ as a hyperplane. Fix a type of totally singular 4-space of $V^{\#}$, and let $F(X)$ be the subspace of that type containing $\langle T(X), I \rangle / \langle I \rangle$. Set $\Sigma = \{F(X) | \langle X \rangle \in \Omega\}$.

THEOREM 6.14. *If* $q \equiv \pmod 3$ *then* Σ *is a spread of* $V^{\#}$ *permuted* 2*-transitively by* $PGU(3, q)$.

Proof. This is proved exactly as in Theorem 6.13. □

The spreads Σ of Theorems 6.13 (i) and 6.14 will be called *unitary spreads*. For $q = 3^{2e+1}$, they are due to Thas [19]. When $q \equiv 2 \pmod 3$, they can be extracted from Tits [21, Ch. IV].

PROPOSITION 6.15. (i) *If* $q = 2$ *then* $\Gamma O^+(8, 2)_\Sigma = A_9$.

(ii) *If* $q = 3$ *then* $\Gamma O^+(8, 3)_\Sigma$ *is isomorphic to the derived group of the Weyl group of type* E_7.

(iii) *If* $q > 3$, *and* $q \equiv 0$ *or* $2 \pmod 3$, *then* $\Gamma O^+(8, q)_\Sigma \trianglerighteq PGU(3, q)$.

(iv) *If* q *is even and* $q > 2$ *then* Σ *is nondesarguesian.*

If $q = 2$ this is well known. The case $q = 3$ is discussed (implicitly) in Kantor and Liebler [11, (2.16), Case 8]. When q is even and $q > 2$, Proposition 6.15 can be proved exactly as in Cohen and Wilbrink [2]. The case of odd $q > 3$ can be handled either geometrically or group theoretically; we omit the details.

For future reference, we will introduce some additional notation concerning the symbols $[a, b; \alpha]$ and $[\gamma, \delta]$ defined in (6.10). Let q be even and abbreviate (cf. § 7)

$$(\sigma; \bar{\sigma}) = [0, 0; \sigma] + \langle N \rangle. \tag{6.16}$$

If $A = \left(\begin{smallmatrix} A_1 & 0 \\ 0 & 1 \end{smallmatrix}\right) \in G$, let $A_1^{(2)}$ be obtained from A_1 by squaring all entries. Then

$$A^{-1}(\sigma; \bar{\sigma})A = (\sigma; \bar{\sigma})A_1^{(2)}(\det A_1)^{-1}, \qquad A^{-1}[\gamma, \delta]A = [\gamma, \delta]A_1, \tag{6.17}$$

where the right sides should be viewed as matrix products.

7. Translation planes produced by unitary spreads. Define V, G and Σ as in § 6, using $q = 2^{2e+1} \equiv 2 \pmod 3$. According to § 3, Σ determines various translation planes $\mathscr{A}((y^{\perp} \cap \Sigma)/y)$ as y ranges over the nonsingular points of V. We will only mention three types of planes.

Probably the most interesting example arises when y contains

$$N = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Note that $Q(N) = 1$. The stabilizer $G_{\langle N \rangle}$ of $\langle N \rangle$ consists of all unitary matrices of the form $\left(\begin{smallmatrix} A_1 & 0 \\ 0 & \delta \end{smallmatrix}\right)$, where $A_1^{-1} = \bar{A}_1^t$ and $\delta\bar{\delta} = 1$. Since this group contains all scalar matrices δI with $\delta\bar{\delta} = 1$, and each such matrix acts trivially on V, we can restrict our attention to the group $GU(2, q)$ of those matrices having $\delta = 1$. Here (Dieudonné [6, p. 46]),

$$GU(2, q) \cong SU(2, q) \times \mathbb{Z}_{q+1} \cong SL(2, q) \times \mathbb{Z}_{q+1}.$$

THEOREM 7.1. *Let $q = 2^{2e+1} > 2$, and set $\mathscr{A} = \mathscr{A}((N^{\perp} \cap \Sigma)/\langle N \rangle)$.*

(i) *$\mathscr{A}$ is a nondesarguesian translation plane of order q^3.*

(ii) $(\mathrm{Aut}\, \mathscr{A})_0 \trianglerighteq SL(2, q)$.

(iii) *$SL(2, q)$ has two orbits on the line L_∞ at infinity, of lengths $q+1$ and $q^3 - q$.*

(iv) *There is a desarguesian subplane $\mathscr{A}_0$ containing 0 and the orbit on L_∞ of length $q+1$, and* Aut $\mathscr{A}$ *induces $\Gamma L(2, q)$ on $\mathscr{A}_0$.*

(v) *There is a cyclic collineation group of order $q+1$ fixing $\mathscr{A}_0$ pointwise.*

Proof. The group D of diagonal matrices diag $(\delta, \delta, 1)$ with $\delta\bar{\delta} = 1$ is a cyclic group of order $q+1$ fixing all q^2 of the vectors (6.16). Since $q > 2$, the desarguesian plane $AG(2, q^3)$ admits no such group. This proves (i), and provides us with a desarguesian subplane $\mathscr{A}_0$ of order q such that (v) holds. Also, (6.17) yields (iii) and (iv).

It remains to prove (ii). This is immediate by Corollary 3.7 and Proposition 6.15 (iii) but can also be proved as follows. Let H be the subgroup of $(\mathrm{Aut}\, \mathscr{A})_0$ generated by all elations, and assume that $H \neq SL(2, q)$. Then H merges the two L_∞ orbits of $SL(2, q)$, and hence is transitive on L_∞. Then $\mathscr{A}$ is desarguesian (Lüneburg [13, pp. 178–179]). This contradiction completes the proof. □

COROLLARY 7.2. *There is a primitive cube root of unity ω such that $(N^{\perp} \cap \Sigma)/\langle N \rangle$ consists of the subspaces (cf. (6.10), (6.16))*

$$\{(a\sigma^2; a\bar{\sigma}^2) + [\gamma\sigma, \gamma\bar{\sigma}] \mid a \in K, \gamma \in F\}, \tag{7.3}$$

$$\{(\gamma; \bar{\gamma})A_1^{(2)} + [a\omega, \bar{\gamma}]A_1 \mid a \in K, \gamma \in F\}, \tag{7.4}$$

where $\sigma \in F$ and $A_1 \in SU(2, q)$.

Proof. Set

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

so $\langle X \rangle \in \Omega$. By Lemma 6.12, $T(X)$ consists of all matrices

$$\begin{pmatrix} a & a & \bar{\gamma} \\ a & a & \bar{\gamma} \\ \gamma & \gamma & 0 \end{pmatrix}.$$

Since $F(X)$ is a totally singular 4-space containing $T(X)$, there is a cube root of unity $\omega \neq 1$ such that $F(X)$ consists of all matrices

$$\begin{pmatrix} a & a & \bar{\gamma} \\ a & a & \bar{\gamma} \\ \gamma & \gamma & 0 \end{pmatrix} + \begin{pmatrix} 0 & b\omega & 0 \\ b\bar{\omega} & b & 0 \\ 0 & 0 & b \end{pmatrix}.$$

Then $\langle N, N^{\perp} \cap F(X)\rangle/N$ is just (7.3) with $\sigma = 1$. By (6.17), all the subspaces (7.3) belong to Σ.

If

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

and $X' = P^{-1}XP$, then $F(X')$ consists of all the matrices

$$\begin{pmatrix} a & \bar{\gamma} & a \\ \gamma & 0 & \gamma \\ a & \bar{\gamma} & a \end{pmatrix} + \begin{pmatrix} 0 & 0 & b\omega \\ 0 & b & 0 \\ b\bar{\omega} & 0 & b \end{pmatrix},$$

and $N^{\perp} \cap F(X')$ consists of those with $a = b$. This gives us the matrices $[a, a; \gamma] + [a + a\bar{\omega}, \bar{\gamma}] = [a, a; \gamma] + [a\omega, \bar{\gamma}]$. Modulo $\langle N\rangle$, this is $(\gamma; \bar{\gamma}) + [a\omega, \bar{\gamma}]$. That the subspaces (7.4) consist of all remaining members of $(N^{\perp} \cap \Sigma)/\langle N\rangle$ now follows from (7.1 iii) and (6.17). □

Remark 1. The representations of $SU(2, q)$ on the spaces $(N^{\perp} \cap V_1)/\langle N\rangle$ and V_2 (cf. (6.11)) are related as indicated in (6.17) by the squaring automorphism of $GF(q^2)$.

Remark 2. The group $GU(2, q)$ has a set of $q^2 - q + 1$ subgroups of order $q + 1$ each of which fixes pointwise a desarguesian subplane of order q. No two of these subplanes have any common points on L_∞. (These assertions are proved by considering the action of $PGU(3, q)$ on $PG(2, q^2)$: the required groups of order $q + 1$ are seen to be the homology groups whose axes are fixed by the group in Theorem 7.1 (v).)

We conclude this section with two more slices of Σ.

Example 7.5. Let $q > 2$. Choose $a \neq 0, 1$ in K, and set

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & a+1 \end{pmatrix}.$$

Then $Q(M) = a^2 + a + 1 \neq 0$. Form $\mathscr{A} = \mathscr{A}((M^{\perp} \cap \Sigma)/\langle M\rangle)$. The group $G_{\langle M\rangle}$ induces $H = \mathbb{Z}_{q+1} \times \mathbb{Z}_{q+1}$ on $\mathscr{A}$. H has three cyclic subgroups of order $q + 1$ each of which fixes pointwise a desarguesian subplane of order q. No two of these subplanes have common points at infinity. This plane is neither desarguesian nor one of those discussed earlier, by Corollary 3.6.

Example 7.6. Let $A \in GU(3, q)$ have order $q^3 + 1$. Then $C_V(A)$ is an anisotropic 2-space. Choosing M in $C_V(A)$ produces a plane of order q^3 having a collineation of order $q^2 - q + 1$. By Lemma 3.6, this plane is neither desarguesian nor a type (4.2 IV) cousin of a desarguesian plane.

8. More spreads and planes. For the next construction, take a field $F = GF(q^3)$ of characteristic 2, and let V consist of all quadruples $x = (a, \beta, \gamma, d)$ with $a, d \in K = GF(q)$ and $\beta, \gamma \in F$. Set $Q(x) = ad + T(\beta\gamma)$, where T is the trace map $F \to K$ defined by $T(\beta) = \beta + \beta^q + \beta^{q^2}$. Then Q is a quadratic form on V, and turns V into an $\Omega^+(8, q)$

space. Define

$$\Sigma(\infty)=\{(0,0,\gamma,d)|\alpha\in F, d\in K\},$$
$$\Sigma(t)=\{(a,\beta,at^{q+q^2}+\beta^q t^{q^2}+\beta^{q^2}t^q, T(\beta t^{q+q^2})|a\in K,\beta\in F\}, \tag{8.1}$$

where $t\in F$. Define linear transformations j and $[t]$ by

$$(a,\beta,\gamma,d)^j=(d,\gamma,\beta,a),$$
$$(a,\beta,\gamma,d)^{[t]}=(a,at+\beta,at^{q+q^2}+\beta^q t^{q^2}+\beta^{q^2}t^q+\gamma, at^{1+q+q^2}+T(\beta t^{q+q^2})+T(\gamma t)+d). \tag{8.2}$$

THEOREM 8.3. (i) $\Sigma=\{\Sigma(t)|t\in F\cap\{\infty\}\}$ *is a spread of* V.

(ii) $G=\langle[t], j|t\in F\rangle$ *is a subgroup of* $O^+(8,q)_\Sigma$ *inducing* $PSL(2,q^3)$ *on* Σ, *acting in the usual* 3*-transitive manner.*

(iii) *If* $q>2$ *then* Σ *is neither desarguesian nor unitary.*

Proof. A straightforward calculation shows that j and $[t]$ preserve Q, $\Sigma(\infty)^j=\Sigma(0)$, $\Sigma(0)^j=\Sigma(\infty)$, $\Sigma(t)^j=\Sigma(t^{-1})$ for $t\neq 0,\ \infty$, $\Sigma(\infty)^{[t]}=\Sigma(\infty)$ and $\Sigma(x)^{[t]}=\Sigma(x+t)$ for $x, t\neq\infty$. This proves (ii). Since $\Sigma(\infty)\cap\Sigma(0)=0$, (i) follows from the 2-transitivity of G. Now assume that $q>z$.

Another calculation proves that G acts irreducibly on V, so that Σ cannot be a desarguesian spread by Lemma 4.1. The argument in [2] shows that $\Gamma O^+(8,q)_\Sigma=(SL(2,q^3)\times GF(q)^*)\cdot \mathrm{Aut}\, GF(q^3)$. Thus, Σ also cannot be unitary, and (iii) holds. □

Examples. Let $q>2$. Note that $(a,\beta,\gamma,d)\to(a/\rho^{1+q+q^2}, \beta/\rho,\ \gamma\rho,\ d\rho^{1+q+q^2})$ preserves Q and Σ whenever $\rho\in F^*$. Thus, the stabilizer of $(1,0,0,1)$ in $SL(2,q^3)$ has order $2(q^2+q+1)$. Using $\Omega^+(8,q^6)$, we find that there is also a nonsingular point whose stabilizer has order $2(q^2-q+1)$. Adding orbit lengths, we obtain all $q^3(q^4-1)$ nonsingular points in V (compare Lemma 4.1!).

There are thus just two types of cousins to consider.

One type has its full collineation group inducing a group of order $2(q^2+q+1)|\mathrm{Aut}\,(GF(q^3))|$ on the line at infinity (by Corollary 3.7). There is an orbit there of length 2.

A second cousin has its full collineation group inducing a group of order $2(q^2-q+1)|\mathrm{Aut}\,(GF(q^3))|$ at infinity.

9. One more spread and one more plane. Dye [8, § 4] constructed a curious spread Σ in an $\Omega^+(8,8)$ space V, having the following properties. $\Gamma O^+(8,8)_\Sigma\cong(A_9\times GF(8)^*)\langle\phi\rangle$, where ϕ is the field automorphism of order 3. (In fact, equality holds here although Dye did not prove this.) $\{F\cap C_V(\phi)|F=F^\phi\in\Sigma\}$ is a desarguesian spread in the $\Omega^+(8,2)$ space $C_V(\phi)$. The A_9 prevents the spread Σ from being equivalent to any we have already discussed.

Let y be a nonsingular point fixed by ϕ. Then there is a subgroup $SL(2,8)$ of A_9 fixing y. This group has two orbits on Σ, of lengths 9 and $\binom{9}{3}6$ (as can be seen from Dye's construction).

THEOREM 9.1. (i) $\mathcal{A}((y^\perp\cap\Sigma)/y)=\mathcal{A}$ *is a nondesarguesian translation plane of order* 8^3.

(ii) $(\mathrm{Aut}\,\mathcal{A})_0$ *has a normal subgroup* $SL(2,8)$ *whose orbits at infinity have lengths* $8+1$ *and* 8^3-8. *(Aut* $\mathcal{A}_0$*) fixes a desarguesian subplane of order* 8.

(iii) $(\mathrm{Aut}\,\mathcal{A})_0$ *contains* $P\Gamma L(2,8)\times\mathbb{Z}_3$.

(iv) $\mathcal{A}$ *is not isomorphic to the plane of order* 8^3 *appearing in Theorem* 7.1.

Proof. (i) and (iv) follow from Corollary 3.6. Now (ii) is proved as in Theorem 7.1(ii), and (iii) follows from the fact that $(A_9)_y=P\Gamma L(2,8)$. □

10. New Kerdock sets. In view of (5.3), each construction of a spread in an $\Omega^+(8, q)$ space produces at least one type of Kerdock set. These can be summarized as follows, in the context of 4×4 skew-symmetric matrices over $GF(q)$.

Example 10.1. Desarguesian Kerdock sets for q even, due to Kerdock [11] and Delsarte and Goethals [4]. Automorphism group: 2-transitive, consisting of all $x \to ax^\sigma + b$ over $GF(q^3)$ if $q > 2$; A_8 if $q = 2$ (see Cameron and Seidel [1]).

Example 10.2. Unitary Kerdock sets via § 6, with $q \equiv 0$ or 2 (mod 3), $q > 2$. Automorphism group: transitive, of order $q^3(q^2-1)\log_p q$ if $q > 3$ where p is the prime dividing q; $O(5, 3)$ if $q = 3$ (Patterson [15]).

Example 10.3. Kerdock sets via § 8, with q even, $q > 2$. Automorphism group: same as (10.1).

Example 10.4. Two Kerdock sets for $q = 8$, via § 9, both with intransitive automorphism groups. (These arise from the two different orbits of $\Gamma O^+(8, 8)_\Sigma$ on Σ.)

When $q = 3$, Example 10.2 is not new. It was discovered by Patterson [15], who gave a very different description, found its automorphism group, and noted that there is just one Kerdock set of 4×4 skew-symmetric matrices over $GF(3)$ (up to equivalence). The $GF(2)$ example is also unique. These $GF(2)$ and $GF(3)$ examples are unusually homogeneous. They can be regarded as cohomological curiosities. Namely, they (or sets of points equivalent to them under triality, cf. § 12) arise because of the exceptional behavior of corresponding first cohomology groups (see Kantor and Liebler [11, (2.16)], where the sets Ω in their Cases 2 and 8 correspond to spreads over $GF(2)$ and $GF(3)$; and also Cameron and Seidel [1]).

If $q = 2$, Kerdock sets produce Kerdock codes as explained in Cameron and Seidel [1]. Inequivalent Kerdock sets yield inequivalent codes.

If q is odd, Kerdock sets have not yet been found to produce interesting error-correcting codes.

There is one further known class of spreads we have not mentioned. In an $\Omega^+(8, 3^{2e+1})$ space there is a spread Σ arising from the Ree group $R(q)$ (cf. Tits [23]). When $e = 0$, this is the spread in (6.15). When $e > 0$, it is a different spread (by (6.15)). These yield the following Kerdock sets.

Example 10.5. Kerdock sets for $q = 3^{2e+1} > 3$. Automorphism group transitive, of order $q^3(q-1)(2e+1)$.

11. Some perfect 1-codes. If V_1 is a 7-dimensional orthogonal space over $GF(q)$, form the graph whose vertices are the totally singular 3-spaces, joining two of them when their intersection has dimension 2. Then a spread of V_1 consists of $q^3 + 1$ vertices such that every other vertex is joined to exactly one of them. Thus, spreads produce perfect 1-codes in this graph. Analogous statements hold for the graph obtained in the same manner from totally isotropic 3-spaces of a symplectic 6-space (Thas [18]; Stanton [16]).

Thus our constructions produce examples of such perfect codes—in fact, large numbers of examples, using $y^\perp \cap \Sigma$ for any nonsingular point y and any spread Σ in §§ 6, 8 or 9. Even the desarguesian spread produces many sections $y^\perp \cap \Sigma$ of this type. Since Σ is uniquely recoverable from $y^\perp \cap \Sigma$ as in § 3, different $\Gamma O^+(8, q)_\Sigma$ orbits of nonsingular points produce inequivalent perfect codes.

Similar statements hold for the symplectic case when q is even, using $(y^\perp \cap \Sigma)/y$ as in § 3. Stanton [17] has recently constructed perfect 1-codes in the space of 3×3 symmetric matrices which are closely related to desarguesian spreads, and which have analogues for every 6-dimensional symplectic spread constructed here.

12. Ovoids. An ovoid in an orthogonal vector space is a set Ω of singular points such that each maximal totally singular subspace meets Ω exactly once. These are much rarer than spreads: Thas [18] has shown that they cannot exist in various orthogonal spaces. However, they can occur in $\Omega^+(8, q)$ spaces. Namely, the image of a spread under a suitable triality map (Tits [22]) is an ovoid, and vice versa. The group of the spread is sent to the group preserving the ovoid.

Thus, ovoids exist in $\Omega^+(8, q)$ spaces for q even or $q \equiv 0$ or $2 \pmod 3$, in view of §§ 4, 6. Examples presumably exist for $q \equiv 1 \pmod 6$, but none are known. No examples are known in dimension larger than 8.

The $\Omega^+(8, 3)$ ovoid is especially pleasant: it is related to the root system of type E_7 (Kantor and Liebler [11, (2.16), Case 8]). The examples in Thas [20] for $\Omega^+(8, 3^{2e+1})$ are among those arising from § 6 and (10.5).

An ovoid of a 4-dimensional unitary space is a set Ω of q^3+1 isotropic points such that each totally isotropic line meets Ω exactly once. The obvious example has Ω consisting of all points of a nonsingular 3-space. Since the generalized quadrangles for $SU(4, q)$ and $\Omega^-(6, q)$ are duals of one another, an ovoid flips to a spread of totally singular lines of an $\Omega^-(6, q)$ space. Such spreads are easy to construct (Dillon [7], Dye [8], Thas [19]). Namely, if S is a 6-space of type $\Omega^-(6, q)$ in an $\Omega^+(8, q)$ space V and if Σ is a spread in V, then $\Sigma \cap S$ is a spread in S. Unfortunately, it is not clear how to recover Σ from $\Sigma \cap S$, so it seems difficult to decide whether or not the large number of resulting $\Omega^-(6, q)$ spreads are equivalent; presumably, $\Omega^-(6, q)$ equivalence implies $\Omega^+(8, q)$ equivalence.

REFERENCES

[1] P. J. CAMERON AND J. J. SEIDEL, *Quadratic forms over GF(2)*, Kon. Ned. Ak. Wet., Proc. A 76 (1973), pp. 1–8.

[2] A. M. COHEN AND H. A. WILBRINK, *The stabilizer of Dye's spread on a hyperbolic quadric in PG(4n − 1, 2) within the orthogonal group*, Rend. Acc. Naz. Lincei (8) 69 (1980), pp. 22–25.

[3] F. DECLERCK, R. H. DYE AND J. A. THAS, *An infinite class of partial geometries associated with the hyperbolic quadric in PG(4n − 1, 2)*, Europ. J. Combinatorics, 1 (1980), pp. 323–326.

[4] P. DELSARTE AND J. M. GOETHALS, *Alternating bilinear forms over GF(q)*, J. Combin. Theory A, 19 (1975), pp. 26–50.

[5] P. DEMBOWSKI, *Finite Geometries*, Springer, Berlin–Göttingen–Heidelberg, 1968.

[6] J. DIEUDONNÉ, *La géométrie des groupes classiques*, Springer, Berlin–Göttingen–Heidelberg 1971.

[7] J. F. DILLON, *On Pall partitions for quadratic forms*, unpublished manuscript 1974.

[7a] ———, *Elementary Hadamard difference sets*, Ph.D. thesis, Univ. of Maryland, College Park, 1974.

[8] R. H. DYE, *Partitions and their stabilizers for line complexes and quadrics*, Ann. Mat. (4), 114 (1977), pp. 173–194.

[9] J. M. GOETHALS, *Nonlinear codes defined by quadratic forms over GF(2)*, Inform. and Control, 31 (1976), pp. 43–74.

[10] W. HAEMERS AND J. H. VAN LINT, *A partial geometry pg(9, 8, 4)*, Combinatorica, to appear.

[11] W. M. KANTOR AND R. A. LIEBLER, *The rank 3 permutation representations of the finite classical groups*, Trans. Amer. Math. Soc., to appear.

[12] A. M. KERDOCK, *A class of low-rate non-linear binary codes*, Inform. and Control, 20 (1972), pp. 182–187.

[13] H. LÜNEBURG, *Translation Planes*, Springer, New York, 1980.

[14] F. J. MACWILLIAMS AND N. J. A. SLOANE, *The theory of error-correcting codes*, North-Holland, Amsterdam, 1977.

[15] N. J. PATTERSON, *A four-dimensional Kerdock set over GF(3)*, J. Combin. Theory (A) 20 (1976), pp. 365–366.

[16] D. STANTON, *Some q-Krawtchouk polynomials on Chevalley groups*, Amer. J. Math., 102 (1980), pp. 625–662.

[17] ———, *Another infinite family of perfect codes*, to appear.

[18] J. A. THAS, *Two infinite classes of perfect codes in metrically regular graphs*, J. Combin. Theory (B) 23 (1977), pp. 236–238.

[19] J. A. THAS, *Ovoids and spreads of finite classical polar spaces*, Geom. Dedicata, 10 (1981), pi. 135–144.
[20] ———, *Polar spaces, generalized hexagons and perfect codes*, J. Comb. Theory (A), 29 (1980), pp. 87–93.
[21] ———, *Some results on quadrics and a new class of partial geometries*, Simon Stevin, to appear.
[22] J. TITS, *Sur la trialité et certains groupes qui s'en déduisent*, Publ. Math. IHES, 2 (1959), pp. 14–60.
[23] ———, *Les groupes simples de Suzuki et de Ree*, Sém. Bourbaki, No. 210, 1960–61.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 2, June 1982

0196-5212/82/0302-0003 $01.00/0

BOUNDS ON THE RELIABILITY POLYNOMIAL FOR SHELLABLE INDEPENDENCE SYSTEMS*

MICHAEL O. BALL† AND J. SCOTT PROVAN‡

Abstract. The reliability polynomial associated with an independence system is $g(p)=\sum_{k=0}^{n} f_k p^k (1-p)^{n-k}$, where f_k is the number of independent sets of cardinality k and n is the cardinality of the ground set. An independence system (T, Γ) is shellable if all maximal independent sets have the same cardinality and if there exists an ordered partition of the set of independent sets into intervals $\{[F_i, G_i]\}_{i=1}^{I}$ (an interval $[F, G]=\{F': F\subseteq F'\subseteq G\}$) where for all n', $n'\leqq I$, $G_{n'}$ is a maximal independent set and $(T, \cup_{i=1}^{n'}[F_i, G_i])$ is an independence system. For the class of shellable independence systems, tight upper and lower bounds are given on $g(p)$, when the number of maximal independent sets and the number of minimum cardinality dependent sets are fixed. These results can be applied to obtain bounds on the reachability measure, which is the probability that a stochastic network (directed or undirected) contains a path from a specified node to all other nodes.

1. Introduction. The problem of computing the reliability of a stochastic network has received significant attention in recent years. All known algorithms (see, for example, [3], [12], [25]) for computing network reliability exactly have running times that grow exponentially with the size of the network. In addition, virtually all network reliability analysis problems of practical interest are known to be NP-hard [2], [20], [24], [28]. Naturally, this fact has led researchers to look toward approximation procedures. In this paper we derive bounds on the reliability polynomial for certain classes of stochastic coherent binary systems. These bounds apply to an important network problem that arises in the design of communications networks, namely, the reachability problem. The reachability problem is that of computing the probability that a stochastic network contains an operating path from a specified node to all other nodes. In the undirected case, it is the problem of computing the probability that a stochastic network is connected. Computation of these bounds requires significantly less effort than exact algorithms.

First we define the general reliability analysis problem. Let T be a finite, nonempty set, Γ a collection of subsets of T and p a real number, $0<p<1$. A *stochastic binary system* is a triple (T, Γ, p) with the following interpretation. The index set T represents a set of components, each of which can be in one of two states: operative or failed. The state of each component is a random event that is independent of the state of any other components. Each component fails with probability p and operates with probability $1-p$. The system can be in either of two states: operative or failed. Its state is a function of the states of the components. We represent the state of the components by the set F of failed components. In particular, F represents the random event in which the components of F are failed and the components of $T-F$ are operative. The probability of the event corresponding to F is $p^{|F|}(1-p)^{|T-F|}$. We let Γ be the family of subsets F of T such that if the components in F fail and the components in $T-F$ operate then the system operates. The probability that the system operates, denoted by $P(\Gamma, p)$, is the probability of occurrence of some event corresponding to an element of Γ. $P(\Gamma, p)$ may be written as a polynomial

$$g_\Gamma(p)=\sum_{i=0}^{n} f_i p^i (1-p)^{n-i}, \tag{1}$$

* Received by the editors April 18, 1980, and in revised form June 15, 1981.

† College of Business and Management, University of Maryland, College Park, Maryland 20742.

‡ Department of Applied Mathematics, State University of New York at Stony Brook, Stony Brook, New York 11790.

where

$$n = |T|,$$

$$f_i = |\{F \in \Gamma: |F| = i\}| \quad \text{for } i = 0, 1, \cdots, n.$$

Note that $0 \leqq f_i \leqq \binom{n}{i}$ for all i. A stochastic binary system is called *coherent* [9] if $\phi \in \Gamma$, $T \notin \Gamma$ and $F \in \Gamma$, $F' \subset F$ imply $F' \in \Gamma$. The *reliability analysis problem* that we consider is to determine $g_\Gamma(p)$ for a stochastic coherent binary system. For any stochastic coherent binary system, we call the ordered pair (T, Γ) an *independence system*. Denote by d the cardinality of a maximum cardinality member of Γ. We have for $i > d$, $f_i = 0$, since no set of cardinality i is contained in Γ. A *circuit* of (T, Γ) is a minimal set F not contained in Γ. Denote by c the cardinality of a minimum cardinality circuit. For $i < c$, $f_i = \binom{n}{i}$ since every set of cardinality i is contained in Γ.

Our approach to generating bounds is similar to the approach suggested by Van Slyke and Frank [29]. They note that f_d and f_c can usually be easily computed ("easily" will be defined more precisely later) and that bounds on the f_i directly produce bounds on g_Γ. In particular, they apply results due to Kruskal [17] and Katona [16] to give an upper bound on g_Γ given f_c and a lower bound on g_Γ given f_d. While our bounds apply to a more restricted class of independence systems, in general they are much tighter since they simultaneously take into account both f_c and f_d and since they take into account special problem structure.

The main results of our paper are tight upper and lower bounds on $g_\Gamma(p)$ given f_d and f_c for a special class of independence systems called *shellable* independence systems. That is, we define polynomials $\underline{g}(p)$ and $\bar{g}(p)$ which satisfy $\underline{g}(p) \leqq g_\Gamma(p) \leqq \bar{g}(p)$ for all p and for which there exists a shellable independence system with reliability polynomials $\underline{g}(p)$ and $\bar{g}(p)$ and with $f_0, \cdots, f_c$ and f_d as given. Our results use a second form of the reliability polynomial given by

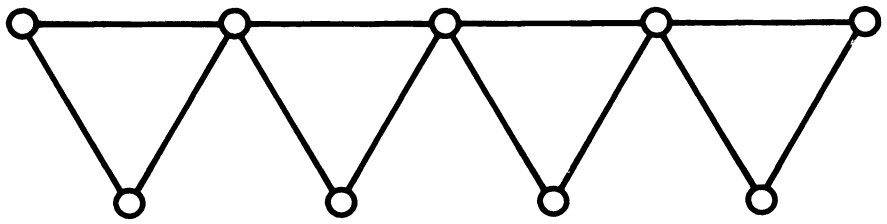
$$g_\Gamma(p) = (1-p)^{n-d} \sum_{i=0}^{d} h_i p^i. \tag{2}$$

For shellable independence systems, we give a particular interpretation to the h_i's which allows us to use results due to Stanley [27] to generate bounds.

In § 2 we give several bounds on the reliability polynomial. First we summarize the Kruskal/Katona results as applied to the reliability polynomial. Then we state our main results which are tight bounds for shellable independence systems.

In § 3 we describe two shellable independence systems arising from graphs. In the first, T is the set of edges in an undirected graph. The system operates if there is an operating path between every pair of nodes. Thus, Γ contains all $F \subset T$ such that $T - F$ contains a spanning tree. A maximal independent set is the complement of a spanning tree and a circuit is a network cut. Figure 1 illustrates the concepts defined earlier for this graph problem. For the second graph problem, T is the set of edges in a directed graph. The system operates if there is an operating directed path from a specified node s to all other nodes. Thus, Γ contains all $F \subset T$ such that $T - F$ contains a spanning arborescence rooted at s. A maximal independent set is the complement of a spanning arborescence, and a circuit is an s-directed cut. Both of these systems arise frequently in the study of communications networks. Throughout the paper, when discussing these graphs, we will assume that the undirected graph contains a spanning tree and that the directed graph contains a spanning arborescence.

In § 4 we derive several properties of the reliability polynomial that are later used in determining the bounds. In particular, it is shown that the coefficients h_i defined by (2) are identical to a set of similar coefficients defined by Stanley [27].



n = number of edges = 12,
d = number of edges − number of nodes + 1 = 4,
c = cardinality of minimum cut = 2,
f_d = number of spanning trees = $81 = \binom{8}{4} + \binom{5}{3} + \binom{2}{2}$,
$f_c = \binom{n}{c}$ − number of minimum cardinality cuts = $54 = \binom{10}{2} + \binom{9}{1}$,
$h_c = \binom{n-d+c-1}{c}$ − number of minimum cardinality cuts = $24 = \binom{7}{2} + \binom{4}{1}$,
$g_\Gamma(p) = (1-p)^{12} + 12p(1-p)^{11} + 54p^2(1-p)^{10} + 108p^3(1-p)^9 + 81p^4(1-p)^8$
$= (1-p)^8(1+8p+24p^2+32p^3+16p^4)$.

FIG. 1

This equivalence enables us to use a powerful result due to Stanley in generating our bounds. Many of the properties contained in § 4 are of interest in their own right and should be useful in other applications. Section 5 contains the proofs of the results contained in § 2.

2. Summary of results. Our approach to generating bounds is to use quantities that, in practical situations, are known a priori or that can be computed efficiently. Many previously developed bounding techniques [14], [19] in network reliability start out by enumerating all maximal independent sets or all circuits. This task in itself can be quite formidable. Our bounding results use one or more of n, d, c, f_d and f_c. For the two graph problems described in § 1, n equals the number of edges and d equals the number of edges minus the number of nodes plus one (see Fig. 1). The coefficient f_d for the undirected problem is the number of spanning trees, and for the directed problem, f_d is the number of spanning arborescences. Both of these quantities can be computed in polynomial time using the matrix tree theorem [7] and an efficient algorithm for computing the determinant of a matrix [1]. For the undirected graph problem, c is the cardinality of a minimum cardinality cut, and for the directed problem, c is the cardinality of a minimum cardinality s-directed cut. These quantities can be computed in polynomial time using network flow techniques [15]. In the undirected case, $f_c \leqq \binom{n}{2}$, [10], and consequently, it can be computed in polynomial time [6]. In the directed case, however, the computation of f_c is known to be NP-hard [20]. However, as Van Slyke and Frank have observed, c is typically small and it is feasible to compute f_c by enumeration. We should note that for planar graphs f_c, in both the directed and undirected cases, can be computed in polynomial time.

Once n, d, c, f_d and f_c are obtained, all subsequent calculations are bounded by a polynomial in n. Thus, in all cases they will lead to computationally efficient bounding procedures. Computational considerations are discussed in [6].

The first bounds we discuss use the results of Kruskal and Katona. In this case we use the first form of the polynomial given by (1). To bound the polynomial (1), we look for vectors of coefficients $(\underline{f}_0, \underline{f}_1, \cdots, \underline{f}_n)$ and $(\bar{f}_0, \bar{f}_1, \cdots, \bar{f}_n)$, with $\underline{f}_i \leqq f_i \leqq \bar{f}_i$ for all i. Since, for $0 < p < 1$, $p^i(1-p)^{n-i} \geqq 0$, such vectors lead directly to bounds on the polynomial. Given the value f_k for some k, Kruskal and Katona were able to give values for $\underline{f}_i$ for $i \leqq k$ and $\bar{f}_i$ for $i \geqq k$ and produce systems that satisfy both simultaneously. The values can be calculated as follows: for any integers m and

k, m can be written uniquely in the *k-canonical form* as

$$m = \binom{m_k}{k} + \binom{m_{k-1}}{k-1} + \cdots + \binom{m_l}{l},$$

where $m_k > m_{k-1} > \cdots > m_l \geqq l \geqq 1$, by choosing $m_k, m_{k-1}, \cdots$ successively satisfying

$$m_i = \max\left\{x : \binom{x}{i} \leqq m - \sum_{j=i+1}^{k} \binom{m_j}{j}\right\}$$

and noting that $m - \sum_{j=i+1}^{k} \binom{m_j}{j} < \binom{m_{i+1}}{i}$ so that $m_k > m_{k-1} > \cdots > m_l \geqq l \geqq 1$. The sequence $(m_k, \cdots, m_l)$ is called the *k-canonical vector* for m. Now for $i \geqq 1$ we define the (i, k)th *lower pseudopower* of m to be

$$m^{(i/k)} = \binom{m_k}{i} + \binom{m_{k-1}}{i-1} + \cdots + \binom{m_l}{i-k+l},$$

where $\binom{m_i}{q} = 0$ if $q < 0$ or $m_i < q$. The bounds given by Kruskal and Katona are $\underline{f}_i = f_k^{(i/k)}$ for $i \leqq k$ and $\bar{f}_i = f_k^{(i/k)}$ for $i \geqq k$. We extend these bounds to bounds on the polynomial by setting $\bar{f}_i = \binom{n}{i}$ for $i < k$ and $\underline{f}_i = 0$ for $i > k$. Thus, we have:

THEOREM 1. *If (T, Γ) is an independence system with $|T| = n$ and $f_k = m$ known, then for $0 < p < 1$*

$$\underline{g}_1(p) \leqq g_\Gamma(p) \leqq \bar{g}_1(p),$$

where

$$\underline{g}_1(p) = \sum_{i=0}^{k-1} m^{(i/k)} p^i (1-p)^{n-i} + m p^k (1-p)^{n-k},$$

$$\bar{g}_1(p) = \sum_{i=1}^{k-1} \binom{n}{i} p^i (1-p)^{n-i} + m p^k (1-p)^{n-k} + \sum_{i=k+1}^{d} m^{(i/k)} p^i (1-p)^{n-i}.$$

Furthermore, these bounds are tight.

The value of k for which f_k is known in Theorem 1 might be c or d or any other index for which f_k was easy to compute. With $k = c = 2$ for the system illustrated in Fig. 1, the Theorem 1 bounds would be computed as follows. The 2-canonical form of $f_c = 54$ is $\binom{10}{2} + \binom{9}{1}$ and $54^{(i/2)} = \binom{10}{i} + \binom{9}{i-1}$. Thus, we have $\underline{g}_1(p) = (1-p)^{12} + 11(1-p)^{11}p + 54(1-p)^{10}p^2$ and $\bar{g}_1(p) = (1-p)^{12} + 12(1-p)^{11}p + 54(1-p)^{10}p^2 + 156(1-p)^9 p^3 + 294(1-p)^8 p^4$.

Theorem 2 gives an extension of Theorem 1 to the commonly encountered case where $f_1, \cdots, f_k$ and f_d are known.

THEOREM 2. *Let (T, Γ) be a rank d independence system with $|T| = n$. Suppose we are given f_d and f_i for $i = 0, 1, \cdots, k$. Then for $0 < p < 1$*

$$\underline{g}_2(p) \leqq g_\Gamma(p) \leqq \bar{g}_2(p),$$

where

$$\underline{g}_2(p) = \sum_{i=0}^{k} f_i p^i (1-p)^{n-i} + \sum_{i=k+1}^{d-1} f_d^{(i/d)} p^i (1-p)^{n-i} + f_d p^d (1-p)^{n-d},$$

$$\bar{g}_2(p) = \sum_{i=0}^{k} f_i p^i (1-p)^{n-i} + \sum_{i=k+1}^{d-1} f_k^{(i/k)} p^i (1-p)^{n-i} + f_d p^d (1-p)^{n-d}.$$

For the system given in Fig. 1, we have $f_4 = 81 = \binom{8}{4} + \binom{5}{3} + \binom{2}{2}$. Thus, with $k = c = 2$ the Theorem 2 bounds for the system are: $g_2(p) = (1-p)^{12} + 12(1-p)^{11}p + 54(1-p)^{10}p^2 + 68(1-p)^9 p^3 + 81(1-p)^8 p^4$ and $\bar{g}_2(p) = (1-p)^{12} + 12(1-p)^{11}p + 54(1-p)^{10}p^2 + 156(1-p)^9 p^3 + 81(1-p)^8 p^4$.

Van Slyke and Frank note that the bounds given by Theorem 1 are not tight for graphs, nor are they tight for matroids. We direct our attention to the more restricted class of shellable independence systems and derive stronger bounds that are tight over this class.

Let (T, Γ) be an independence system in which all maximal independence sets have the same cardinality. In this case, we call d the *rank* of (T, Γ) and a maximal independent set a *basis*. Shellability can be defined in terms of the existence of a particular partition of Γ. Any partition of Γ generates an expression for $P(\Gamma, p)$ or equivalently for $g_\Gamma(p)$. For any $\Gamma' \subseteq \Gamma$ denote by $P(\Gamma', p)$ the probability of the event corresponding to Γ'. Then, if $\{\Gamma_j\}_{j=1}^J$ is a partition of Γ so that $\Gamma_i \cap \Gamma_j = \varnothing$ for all i and j and $\bigcup_{j=1}^J \Gamma_j = \Gamma$, then

$$P(\Gamma, p) = g_\Gamma(p) = \sum_{j=1}^{J} P(\Gamma_j, p). \tag{3}$$

For any $F, G \subseteq T$ with $F \subseteq G$ define the *interval between F and G* as the family of subsets, $[F, G] = \{F' \subseteq T : F \subseteq F' \subseteq G\}$. Thus, $[F, G]$ represents the event in which all components of F fail, all components of $T - G$ operate and the components of $G - F$ can either operate or fail. It is easy to see that $P([F, G], p) = p^{|F|}(1-p)^{n-|G|}$. A partition $\{\Gamma_j\}_{j=1}^J$ of Γ is called an *interval partition* if $\Gamma_j = [F_j, G_j]$ for all j. Applying (3) gives

$$g_\Gamma(p) = \sum_{j=1}^{J} p^{|F_j|}(1-p)^{n-|G_j|}. \tag{4}$$

We say that (T, Γ) is *partitionable* if there exists an interval partition of Γ, with G_j a basis for all j. For partitionable systems, $n - |G_j| = n - d$; we can factor $(1-p)^{n-d}$ out of (4), and we derive (2), where we define

$$h_i = |\{F_j : |F_j| = i\}|. \tag{5}$$

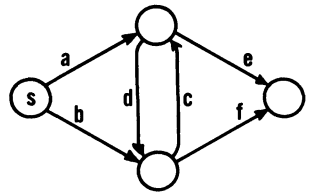
We say that (T, Γ) is *shellable* if there exists an interval partition $\{[F_k, G_k]\}_{k=1}^K$ with G_k a basis for all k and with $(T, \bigcup_{k=1}^j [F_k, G_k])$ an independence system for all j. Partitionable systems have been defined and studied independently by Ball and Nemhauser [5] and by Stanley [27]. Shellable systems have drawn considerable attention using several equivalent versions of the definition given above. (See, for example, [11], [13], [18], [21] and [27].) Figure 2 illustrates an interval partition of an independence system which also satisfies the condition for shellability.

It follows directly from the interpretation given by (5) that, for partitionable systems $h_i \geqq 0$ for all i. Stanley [27, Thm. 6], gives a deep and powerful characterization for the h_i vector for shellable systems, which is restated in Theorem 5 of this paper. It is this characterization which is used to derive bounds on the reliability polynomial. Before stating our bounds we define a second type of pseudopower. For any sequence $(m_k, m_{k-1}, \cdots, m_l)$ of integers $k \geqq l \geqq 1$ and any integer $i \geqq 0$, define the (i, k)th *upper pseudopower of* $(m_k, \cdots, m_l)$ to be

$$(m_k, \cdots, m_l)^{\langle i/k \rangle} = \binom{m_k - k + i}{i} + \binom{m_{k-1} - k + i}{i-1} + \cdots + \binom{m_l - k + i}{i - k + l}.$$

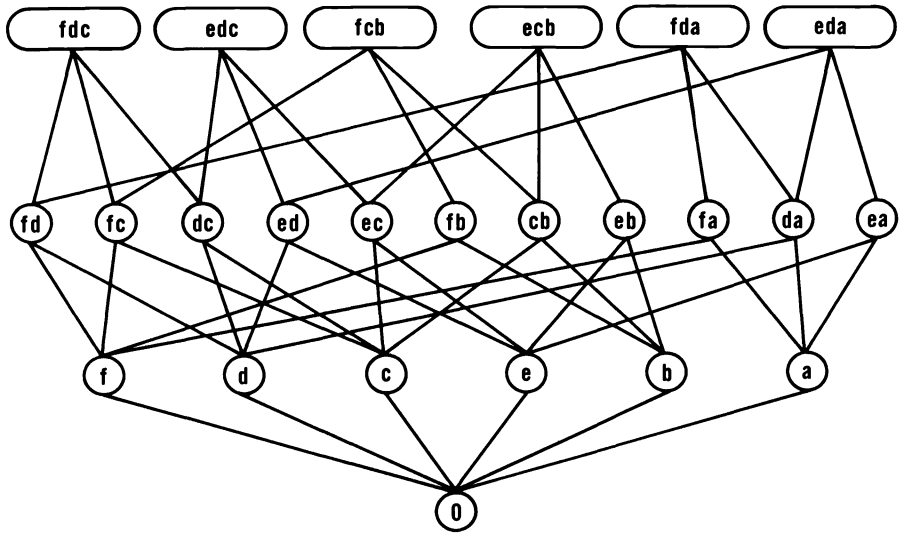
Here again we set $\binom{p}{q} = 0$ in the cases when p, q do not satisfy $p \geqq q \geqq 0$, *except for the special case* $\binom{-1}{0} = 1$. The (i, k)th upper pseudopower of an integer m denoted $m^{\langle i/k \rangle}$ is defined to be the (i, k)th upper pseudopower of its k-cannonical vector. The next two theorems give bounds on the reliability polynomial that are based on Stanley's result. The proofs of the theorems are given in § 5.

Theorem 3 gives bounds on the reliability polynomial for shellable independence systems, given d and h_i for $i \leqq k$. Section 4 gives formulas for computing h_i for $i \leqq k$,



$T = \{a, b, c, d, e, f\}$

$\Gamma = \{S \subseteq T: T - S \text{ contains a spanning arborescence rooted at } s\}$



Ordered Partition:

{[∅, {fdc}], [{e}, {edc}], [{b}, {fcb}], [{eb}, {ecb}],
[{a}, {fda}], [{ea}, {eda}]}

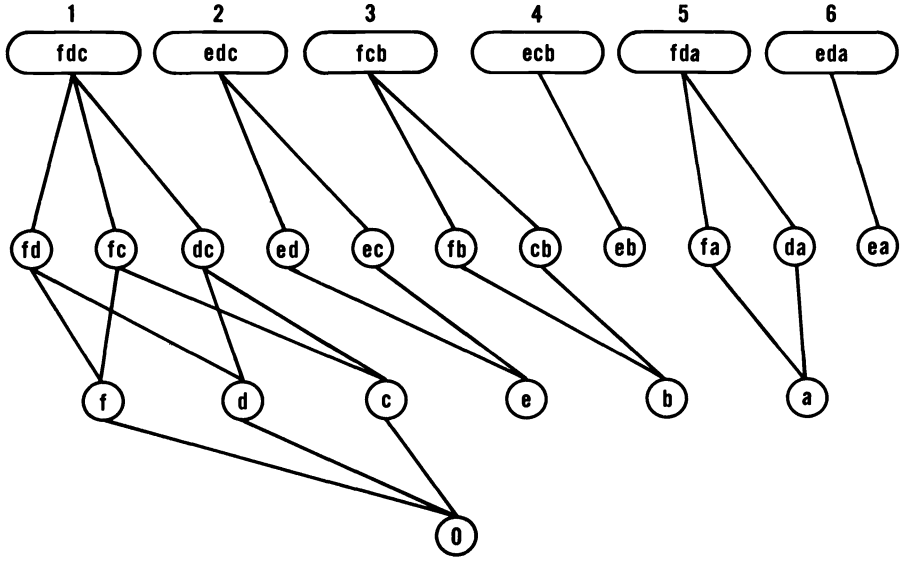


FIG. 2

given f_i for $i \leqq k$. As before, a likely candidate for the value of k is c; this case was studied by Bjorner [8] under the assumption that (T, Γ) is a matroid.

THEOREM 3. *Let (T, Γ) be a rank d shellable independence system with $|T| = n$. Suppose we are given h_i for $i = 0, 1, \cdots, k$. Then for $0 < p < 1$*

$$\underline{g}_3(p) \leqq g_\Gamma(p) \leqq \bar{g}_3(p),$$

where

$$\underline{g}_3(p) = (1-p)^{n-d} \sum_{i=0}^{k} h_i p^i,$$

$$\bar{g}_3(p) = (1-p)^{n-d}\left[\sum_{i=0}^{k} h_i p^i + \sum_{i=k+1}^{d} h_k^{\langle i/k \rangle} p^i\right].$$

Further, these bounds are tight for the class of shellable independence systems.

The proof of this result is given in § 5.

With $k = c = 2$, the Theorem 3 lower bound for the system given in Fig. 1 is $\underline{g}_3(p) = (1-p)^8(1+8p+24p^2)$. To compute $\bar{g}_3(p)$ we require $h_2^{\langle i/2 \rangle}$. The 2-canonical form of $h_2 = 24$ is $\binom{7}{2} + \binom{4}{1}$ and $24^{\langle i/2 \rangle} = \binom{5+i}{i} + \binom{2+i}{i-1}$. Thus, we have $\bar{g}_3(p) = (1-p)^8(1+8p+24p^2+66p^3+146p^4)$.

Our final result, Theorem 4, gives a stronger set of bounds for the case where f_d is also known. f_d has a particular significance on the form of the polynomial given by (2) since, for any system, $\sum_{i=0}^{n} h_i = f_d$. (This relation will be proven in § 4.) Before stating Theorem 4, we define some additional notation. For nonnegative integers m, d and k, $d > k$, let the *(k, d)-factor of m* be that number x that solves

$$x - x^{\langle k/d \rangle} = m.$$

The (k, d)-factor of m can be calculated by simultaneously constructing the d-canonical representation of x,

$$\binom{x_d}{d} + \binom{x_{d-1}}{d-1} + \cdots + \binom{x_l}{l},$$

and the expression for the (k, d)th upper pseudopower of x,

$$\binom{x_d - d + k}{k} + \binom{x_{d-1} - d + k}{k-1} + \cdots + \binom{x_l - d + k}{k - d + l},$$

so that their difference equals m. That is, $x_d, x_{d-1}, \cdots$ can be chosen successively as follows. For $i > d - k$,

$$x_i = \max\left\{x : \binom{x}{i} - \binom{x - d + k}{i - d + k} \leqq m - \sum_{j=i+1}^{d}\left[\binom{x_j}{j} - \binom{x_j - d + k}{j - d + k}\right]\right\},$$

and if $r = m - \sum_{j=d-k+1}^{d}\left[\binom{x_j}{j} - \binom{x_j - d + k}{j - d + k}\right] > 0$, then $x_{d-k}, x_{d-k-1}, \cdots, x_l$ comprise the $(d-k)$ canonical vector for r. It is easy to show that this is a valid process in that $\binom{x_d}{d} + \binom{x_{d-1}}{d-1} + \cdots + \binom{x_l}{l}$ is, in fact, the d-canonical representation of x.

We can now state our main theorem.

THEOREM 4. *Let (T, Γ) be a rank d shellable independence system with $|T| = n$. Suppose we are given f_d and h_i for $i = 0, 1, \cdots, k$. Define the integer $\bar{r}$ and vector $\underline{m}$ as follows:*

$$\bar{r} = \max\left\{r : \sum_{i=k+1}^{r} h_k^{\langle i/k \rangle} \leqq f_d - \sum_{j=0}^{k} h_j\right\},$$

$$\underline{m} = (m_d - 1, \cdots, m_l - 1),$$

where $(m_d, \cdots, m_l)$ is the d-canonical vector for the (k, d)-factor of $f_d - \sum_{j=0}^{k} h_j$.

Then for $0<p<1$

$$\underline{g}_4(p) \leqq g_\Gamma(p) \leqq \bar{g}_4(p),$$

where

$$\underline{g}_4(p) = (1-p)^{n-d}\left(\sum_{i=0}^{k} h_i p^i + \sum_{i=k+1}^{d} \underline{m}^{\langle i/d\rangle} p^i\right),$$

$$\bar{g}_4(p) = (1-p)^{n-d}\left(\sum_{i=0}^{k} h_i p^i + \sum_{i=k+1}^{\bar{r}} h_k^{\langle i/k\rangle} p^i + \left(f_d - \sum_{i=0}^{k} h_i - \sum_{i=k+1}^{\bar{r}} h_k^{\langle i/k\rangle}\right) p^{\bar{r}+1}\right).$$

Furthermore, these bounds are tight for the class of shellable independence systems.

The proof is given in § 5.

With $k=c=2$, the Theorem 4 bounds for the system given in Fig. 1 can be computed as follows. First, we require $\bar{r}$ and $\underline{m}$: $\bar{r} = \max\{r: \sum_{i=3}^{r} 24^{\langle i/2\rangle} \leqq 81-33\} = 2$. To compute the (2, 4)-factor of 48, we first find $48 = \binom{7}{4} - \binom{5}{2} + \binom{6}{3} - \binom{4}{1} + \binom{4}{2} - \binom{2}{0} + \binom{2}{1}$ so that the (2, 4)-factor of 48 is $\binom{7}{4} + \binom{6}{3} + \binom{4}{2} + \binom{2}{1} = 62$. Thus, $\underline{m} = (6, 5, 3, 1)$. We now have $\underline{g}_4(p) = (1-p)^8(1+8p+24p^2+19p^3+29p^4)$ and $\bar{g}_4(p) = (1-p)^5(1+8p+24p^2+48p^3)$. When converted into the form above, the polynomials derived in Theorem 2 for this system become $\underline{g}_2(p) = (1-p)^8(1+8p+24p^2-8p^3+56p^4)$ and $\bar{g}_2(p) = (1-p)^8(1+8p+24p^2+80p^3-32p^4)$. Both bounds derived in Theorem 4, therefore, are tighter than those derived in Theorem 2.

3. Shellable independence systems arising from graphs. In this section we verify that the two classes of independence systems defined in § 1, namely, the class related to spanning trees of undirected graphs and the class related to spanning arborescences of directed graphs, are shellable.

A. *Spanning tree systems.* It is well known that the independence system defined in § 1 for undirected graphs is a matroid. This matroid is commonly referred to as the bond-matroid or co-graphic matroid. Provan and Billera [21], Corollary 3.3.2] have shown that all matroids are shellable. Thus, it follows that

PROPOSITION 1. *If T is the set of edges in an undirected connected graph and $\Gamma = \{S \subseteq T: T-S$ contains a spanning tree$\}$, then the independence system (T, Γ) is shellable.*

B. *Rooted spanning arborescence systems.* We may represent the independence system, defined for directed graphs in § 1, in the following interesting way. Let the vertices of the directed graph be $\{r, v_1, v_2, \cdots, v_M\}$, with r the root vertex, and let the edges of the directed graph be $\{e_1, e_2, \cdots, e_N\}$. Define the $M \times N$ matrix A to have entries

$$a_{ij} = \begin{cases} 1 & \text{if } e_j = (v_k, v_i) \text{ for some } k, \\ -1 & \text{if } e_j = (v_i, v_k) \text{ for some } k, \\ 0 & \text{otherwise.} \end{cases}$$

Now consider the linear system in nonnegative variables $x = (x_1, x_2, \cdots, x_N)$,

$$P_\Gamma: \quad Ax = \underline{1}, \qquad x \geqq 0,$$

where $\underline{1}$ is a vector of M ones. It is well known that the basic feasible solutions to P_Γ correspond to sets of columns whose edges form a spanning arborescence rooted at r. Further, the system P_Γ is nondegenerate. Provan and Billera [22, § 2a] show that

the shellability of convex polytopes developed by Brugesser and Mani [11] can be extended to cover any independence system derived from a nondegenerate linear system by taking as a basis the complement of the set of indices of positive components for any basic feasible solution.

PROPOSITION 2. *If T is the set of edges in a directed graph that contains a spanning arborescence rooted at r and $\Gamma = \{S \subseteq T: T - S$ contains a spanning arborescence rooted at $r\}$, then the independence system (T, Γ) is shellable.*

4. Properties of the reliability polynomial. Equations (1) and (2) define two forms of the reliability polynomial. In this section we give several properties of the polynomial which are used subsequently in the paper.

Propositions (3) and (4) give formulas for converting between the two forms of the polynomial. The proofs of these propositions are given in [4].

PROPOSITION 3. *For any independence system, if h_j and f_k are defined as in* (1) *and* (2), *then for $0 \leqq j \leqq n$*

$$h_j = \sum_{k=0}^{j} f_k (-1)^{j-k} \binom{d-k}{j-k}. \tag{6}$$

PROPOSITION 4. *For any independence system, if h_j and f_k are defined as in* (1) *and* (2), *then for $0 \leqq i \leqq n$*

$$f_i = \sum_{j=0}^{i} \binom{d-j}{i-j} h_j. \tag{7}$$

COROLLARY 1. *Under the assumptions of Proposition* 4,

$$f_d = \sum_{j=0}^{d} h_j. \tag{8}$$

It is interesting to note that for partitionable systems (8) is trivially true and (7) can be easily derived using geometric arguments.

Equation (6) can be used to show that for $j < c$

$$h_j = \binom{n-d+j-1}{j} \tag{9}$$

and

$$h_c = \binom{n-d+c-1}{c} - \left(\binom{n}{c} - f_c\right). \tag{10}$$

Stanley [27] defines an "h-vector" for general independence systems and gives several properties of it. Proposition 5 shows that the h-vector defined by (2) and (6) is identical to the h-vector defined by Stanley. To avoid confusion, we initially define Stanley's h-vector as a b-vector. Let the Hilbert function for the sequence $f_0, \cdots, f_d$ be defined by

$$H(m) = \begin{cases} 1 & \text{if } m = 0, \\ \sum_{i=0}^{d-1} f_{i+1} \binom{m-1}{i} & \text{if } m > 0. \end{cases}$$

Define the vector $(b_0, b_1, \cdots, b_d)$ as the solution to the equation

$$(1-p)^d \sum_{m=0}^{\infty} H(m) x^m = \sum_{j=0}^{d} b_j x^j. \tag{11}$$

PROPOSITION 5. *For any independence system, if h_j are defined by* (2) *and b_j by* (11), *then for $0 \leqq j \leqq d$, $h_j = b_j$.*

Proof.

$$g_\Gamma(p) = \sum_{k=0}^{d} f_k p^k (1-p)^{n-k} = (1-p)^n \sum_{k=0}^{d} f_k \frac{p^k}{(1-p)^k}$$

$$= (1-p)^n \left(\left[\sum_{k=1}^{d} f_k p^k \sum_{i=0}^{\infty} \binom{k+i-1}{k-1} p^i \right] + 1 \right)$$

$$= (1-p)^n \left(\left[\sum_{k=1}^{d} f_k \sum_{m=0}^{\infty} \binom{m-1}{k-1} p^m \right] + 1 \right)$$

$$= (1-p)^n \left(\sum_{m=0}^{\infty} \left[\sum_{k=1}^{d} f_k \binom{m-1}{k-1} \right] p^m + 1 p^0 \right),$$

where $\binom{n}{m} = 0$ for $m < n$. By definition of $H(m)$, we now have

$$g_\Gamma(p) = (1-p)^n \sum_{m=0}^{\infty} H(m) p^m = (1-p)^{n-d} \sum_{j=0}^{d} b_j p^j.$$

Comparing this expression term for term with (2) gives $b_j = h_j$ for all j. □

We can, therefore, refer to the h-vector of an independence system independently of which context it has been defined.

As was noted in § 2, a sufficient condition for one polynomial of the form given by (1) to dominate another is that $\bar{f}_i \geqq \underline{f}_i$ for all i. Proposition 6 gives similar conditions for polynomials of the form given by (2).

PROPOSITION 6. *Given two vectors of integers $(\bar{h}_0, \bar{h}_1, \cdots, \bar{h}_d)$ and $(\underline{h}_0, \underline{h}_1, \cdots, \underline{h}_d)$, if for all j*

$$\sum_{k=0}^{j} \bar{h}_k \geqq \sum_{k=0}^{j} \underline{h}_k,$$

then $\bar{g}(p) \geqq \underline{g}(p)$ for all $0 < p < 1$, where $\bar{g}(p) = \sum_{k=0}^{d} \bar{h}_k p^k$ and $\underline{g}(p) = \sum_{k=0}^{d} \underline{h}_k p^k$.

The result follows by an inductive argument that uses the fact that p^k is a monotonically nonincreasing function of k.

This section has given several properties of the reliability polynomial. The proofs given in § 5 use these properties. However, we emphasize that this section did not use any properties particular to shellable independence systems. Consequently, many of the results may be useful for deriving bounds for more general classes of independence systems.

5. Proof of main results. In this section we prove our main results, Theorems 3 and 4. These results are based on a deep and powerful characterization of the vector $(h_0, h_1, \cdots, h_d)$ due to Stanley [27]:

THEOREM 5 (Stanley). *Let $(h_0, h_1, \cdots, h_d)$ be an integer vector. Then the following are equivalent*:

(i) *$(h_0, \cdots, h_d)$ is the h-vector for some rank d shellable independence system,*

(ii) *$h_0 = 1$ and $0 \leqq h_{j+1} \leqq h_j^{\langle j+1/j \rangle}$ for $0 \leqq j \leqq d-1$.*

For our purposes, a more useful version of Theorem 5 is given by:

COROLLARY 2. *Let $(h_0, h_1, \cdots, h_d)$ be an integer vector. Then the following are equivalent*:

(i) *$(h_0, \cdots, h_d)$ is the h-vector for some rank d shellable independence system,*

(ii′) *$h_0 = 1$ and $0 \leqq h_j \leqq h_i^{\langle j/i \rangle}$ for $0 \leqq i < j \leqq d$.*

The equivalence of (ii) and (ii′) can be obtained from a simple manipulation of upper pseudopowers. We call any vector that satisfies (ii′) an *O-sequence.*

Theorem 5 and Corollary 2 actually hold for a stronger class of systems called *Cohen–Macaulay complexes,* which are characterized entirely by their global and local homology. Notable examples are simplicial spheres and balls, which include unshellable systems. All known practical classes of Cohen-Macaulay complexes, however, are shellable. In fact, testing shellability seems to be the most tractable way to determine whether a system is Cohen–Macaulay.

Theorem 3 can be easily demonstrated. The lower bound in Theorem 3 and the fact that the bound is tight follow directly from Proposition 6 and Corollary 2. The upper bound also follows directly from these results. The tightness of the bound follows from Corollary 2 by noting that if $h_k \leqq h_i^{\langle k/i \rangle}$ for $i < k$, then $h_k^{\langle j/k \rangle} \leqq h_i^{\langle j/i \rangle}$ for $j > k$. Consequently, if a vector $(h_0, h_1, \cdots, h_d)$ satisfies (ii′) for h_k, it will satisfy (ii′) for h_i for $i < k$.

We now proceed with the proof of Theorem 4. Figure 3 illustrates the idea behind the proof. Since $h_j \geqq 0$ and $\sum_{j=0}^{d} h_j = f_d$, we can interpret assigning values to the h_j's

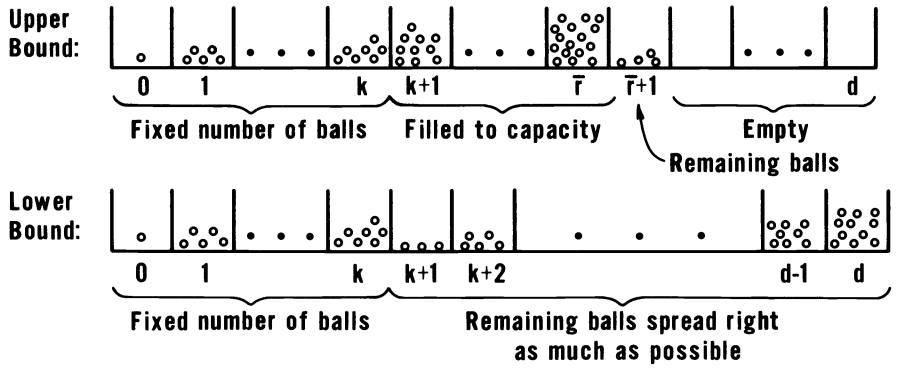


FIG. 3

as placing f_d balls into $d+1$ boxes. We are given the number of balls in boxes 0 through k. Proposition 7 states, essentially, that to obtain an upper bound on g_Γ we would like to place as many balls as possible into the lowest numbered boxes, i.e., $h_k^{\langle k+1/k \rangle}$ into box $k+1$, $h_k^{\langle k+2/k \rangle}$ into box $k+2$, etc., until we run out of balls (the total if f_d) in box $\bar{r}+1$. Again, by Proposition 7, to obtain a lower bound we would like to place as many balls as possible into the highest numbered boxes and as few as possible into the lowest numbered boxes. However, so that (ii′) will be satisfied, in order to place any ball in box d, we must place some in box $d-1$, etc., until we are finally required to place some in box $k+1$. Lemmas 1 through 3 establish that $\underline{m}^{\langle d/d \rangle}$ is the maximum number of balls we can place in box d and that we must put $\underline{m}^{\langle i/d \rangle}$ in box i, $i = d-1, d-2, \cdots, k+1$.

Proposition 7 provides a valuable technique for evaluating bounds on reliability and provides the bounds of Theorem 4 in a less compact form.

PROPOSITION 7. *Let* (T, Γ) *be a rank d shellable independence system with* $|T| = n$. *Suppose we are given* f_d *and* h_i *for* $i = 0, 1, \cdots, k$, *then for* $0 < p < 1$,

$$\underline{g}_4(p) \leqq g_\Gamma(p) \leqq \bar{g}_4(p),$$

where

$$\underline{g}_4(p)=(1-p)^{n-d}\sum_{i=0}^{d}\underline{h}_i p^i, \qquad \bar{g}_4(p)=(1-p)^{n-d}\sum_{i=0}^{d}\bar{h}_i p^i$$

and $\{\underline{h}_i\}_{i=0}^{d}$ *and* $\{\bar{h}_i\}_{i=0}^{d}$ *are computed recursively as follows*:

for $i=0,1,\cdots,k$,

$$h_i=\bar{h}_i=\underline{h}_i \tag{12}$$

and for $i=k+1,k+2,\cdots,d$,

$$\underline{h}_i=\min\left\{m:\sum_{j=0}^{i-1}\underline{h}_j+\sum_{j=1}^{d}m^{\langle j/i\rangle}\geqq f_d\right\}, \tag{13}$$

$$\bar{h}_i=\max\left\{m:m\leqq\bar{h}_{i-1}^{\langle i/i-1\rangle},\ \sum_{j=0}^{i-1}\bar{h}_j+m\leqq f_d\right\}. \tag{14}$$

Proof. It is clear from the definitions of $(\underline{h}_0,\cdots,\underline{h}_d)$ and $(\underline{h}_0,\cdots,\underline{h}_d)$ that they are O-sequences that sum to f_d. Now $g_\Gamma(p)=(1-p)^{n-d}\sum_{i=0}^{d}h_i p^i$, where $(h_0,\cdots,h_d)$ is an O-sequence, and so by Proposition 6 we need only prove that for all $i=0,\cdots,d$.

$$\sum_{j=0}^{i}\underline{h}_j\leqq\sum_{j=0}^{i}h_j\leqq\sum_{j=0}^{i}\bar{h}_j.$$

The right-hand inequality follows immediately from the fact that for some q

$$\bar{h}_j=h_k^{\langle j/k\rangle}\geqq h_j, \qquad j=k+1,\cdots,q-1$$

and

$$\sum_{j=0}^{q}\bar{h}_j=f_d.$$

For the left-hand inequality, we assume the conclusion is false and let q be the smallest integer such that

$$\sum_{j=0}^{q}\underline{h}_j>\sum_{j=0}^{q}h_j.$$

Then $\underline{h}_q>h_q$, and so

$$\sum_{j=0}^{d}h_j\leqq\sum_{j=0}^{q}h_j+\sum_{j=q+1}^{d}h_q^{\langle j/q\rangle}<\sum_{j=0}^{q}\underline{h}_j+\sum_{j=q+1}^{d}(\underline{h}_q-1)^{\langle j/q\rangle}<f_d,$$

a contradiction. This proves the proposition. □

The fact that $\bar{h}$ as defined by Theorem 4 gives the required upper bound follows immediately from Proposition 7 and Corollary 2. To establish the lower bounds for Theorem 4, we first prove three technical lemmas.

LEMMA 1. *For any positive integers* m, i *and* d,

$$m^{\langle i/d\rangle}=\sum_{j=0}^{i}(m_d-1,\cdots,m_l-1)^{\langle j/d\rangle}$$

where $(m_d,\cdots,m_l)$ *is the* d*-canonical vector of* m.

Proof. We use the well-known identity

$$\binom{p+k+1}{k} = \binom{p+k}{k} + \cdots + \binom{p+1}{1} + \binom{p+0}{0},$$

which holds for any integer $p \geqq -1$, with the modifications given earlier. Now

$$\begin{aligned}
&\sum_{j=0}^{i} (m_d - 1, m_{d-1} - 1, \cdots, m_l - 1)^{\langle j/d \rangle} \\
&\quad = \sum_{j=0}^{i} \left[\binom{m_d - 1 + j - d}{j} + \binom{m_{d-1} - 1 + j - d}{j-1} + \cdots + \binom{m_l - 1 + j - d}{l + j - d} \right] \\
&\quad = \sum_{j=0}^{i} \binom{m_d - 1 + j - d}{j} + \sum_{j=1}^{i} \binom{m_{d-1} - 1 + j - d}{j-1} + \cdots + \sum_{j=d-l}^{i} \binom{m_l - 1 + j - d}{l + j - d} \\
&\quad = \sum_{j=0}^{i} \binom{m_d - 1 + j - d}{j} + \sum_{j=0}^{i-1} \binom{m_{d-1} - 1 + j - d + 1}{j} + \cdots + \sum_{j=0}^{i-d+l} \binom{m_l - 1 + j - l}{j} \\
&\quad = \binom{m_d + i - d}{i} + \binom{m_{d-1} + i - d}{i-1} + \cdots + \binom{m_l + i - d}{l + i - d} \\
&\quad = m^{\langle i/d \rangle},
\end{aligned}$$

and this establishes the lemma. □

LEMMA 2. *Let m and d be positive integers, and let m have d-canonical vector $(m_d, \cdots, m_l)$. Then:*

(i) *if $l > 1$, then $m+1$ has d-canonical vector $(m_d, \cdots, m_l, l-1)$;*

(ii) *if $l = 1$, then $m+1$ has d-canonical vector $(m_d, \cdots, m_{q+1}, m_{q+1})$, where q is the smallest index for which $m_{q+1} > m_q + 1$, or if no such q exists, the d-canonical vector is $(m_d + 1)$.*

Proof. Case (i) follows directly from the definitions and uniqueness of the r-canonical representation. Case (ii) can be obtained by application of the identity

$$\binom{m_q + 1}{q} = \binom{m_q}{q} + \binom{m_{q-1}}{q-1} + \cdots + \binom{m_1}{1} + 1,$$

which is given in the proof of Lemma 1. □

LEMMA 3. *Let d and l be integers $d \geqq l \geqq 1$, and let $\underline{m} = (m_d, m_{d-1}, \cdots, m_l)$ be a sequence of integers with $m_d > m_{d-1} > \cdots > m_l \geqq l - 1$. Then, for any integers i and j, $1 \leqq i \leqq j \leqq d$, $(\underline{m}^{\langle i/d \rangle})^{\langle j/i \rangle} \geqq \underline{m}^{\langle j/d \rangle}$.*

Proof. Let p be the smallest index for which $m_p \geqq p$. If $i > d - l$, then

$$m^{\langle i/d \rangle} = \binom{m_d + i - d}{i} + \cdots + \binom{m_l + i - d}{l + i - d} = \binom{m_d + i - d}{i} + \cdots + \binom{m_p + i - d}{p + i - d},$$

the last expression being the k-canonical representation for $\underline{m}^{\langle i/d \rangle}$. Thus,

$$\begin{aligned}
(\underline{m}^{\langle i/d \rangle})^{\langle j/i \rangle} &= \binom{m_d + j - d}{j} + \cdots + \binom{m_p + j - d}{p + j - d} \\
&= \binom{m_d + j - d}{j} + \cdots + \binom{m_l + j - d}{l + j - d} \\
&= \underline{m}^{\langle j/d \rangle},
\end{aligned}$$

since $\binom{m_{s+j-d}}{s+j-d} = 0$ for $s = l, \cdots, p-1$.

If $i \leqq d-l$, then

$$\underline{m}^{\langle i/d\rangle} = \binom{m_d+i-d}{i} + \cdots + \binom{m_{d-i}+i-d}{0}$$

$$= \binom{m_d+i-d}{i} + \cdots + \binom{m_{d+i-1}+i-d}{1} + 1,$$

since $m_{d-1}+i-d \geqq -1$. By Lemma 2, if $p > d-i$, then

$$\underline{m}^{\langle i/d\rangle} = \binom{m_d+i-d}{i} + \cdots + \binom{m_p+i-d}{p+i-d} + \binom{p-1+i-d}{p-1+i-d},$$

and if $p \leqq d-i$, then

$$\underline{m}^{\langle i/d\rangle} = \binom{m_d+i-d}{i} + \cdots + \binom{m_{q+1}+i-d}{q+1+i-d} + \binom{m_q+1+j-d}{q+i-d},$$

where q is the smallest integer greater than $d-i$ for which $m_{q+1} > m_q+1$. In the first case,

$$(\underline{m}^{\langle i/d\rangle})^{\langle j/i\rangle} = \binom{m_d+j-d}{j} + \cdots + \binom{m_p+j-d}{p+j-d} + \binom{p-1+j-d}{p-1+j-d}$$

$$\geqq \binom{m_d+j-d}{j} + \cdots + \binom{m_p+j-d}{p+j-d} + \binom{p-2+j-d}{p-1+j-d} + \cdots + \binom{l-1+j-d}{l+j-d}$$

$$= \underline{m}^{\langle j/d\rangle},$$

and in the second case, using the identity in Lemma 1,

$$(\underline{m}^{\langle i/d\rangle})^{\langle j/i\rangle} = \binom{m_d+j-d}{j} + \cdots + \binom{m_{q+1}+j-d}{q+1+j-d} + \binom{m_q+1+j-d}{q+j-d}$$

$$= \binom{m_d+j-d}{j} + \cdots + \binom{m_{q+1}+j-d}{q+1+j-d}$$

$$+ \binom{m_q+j-d}{q+j-d} + \binom{m_q-1+j-d}{q-1+j-d} + \cdots + \binom{m_q-q+1}{1} + \binom{m_q-q}{0}$$

$$\geqq \binom{m_d+j-d}{j} + \cdots + \binom{m_l+j-d}{l+j-d}$$

$$= \underline{m}^{\langle j/d\rangle},$$

since $m_q - s \geqq m_{q-s}$ for $s = l, \cdots, q$. This completes the proof of the lemma. □

It remains to show that the $\underline{h}$ defined in Theorem 4 for the lower bound satisfies Proposition 7. From Lemma 3, we have for $i = k+1, \cdots, d-1$,

$$\underline{h}_{i+1} = \underline{m}^{\langle i+1/d\rangle} \leqq (\underline{m}^{\langle i/d\rangle})^{\langle i+1/i\rangle} = \underline{h}_i^{\langle i+1/i\rangle}$$

and from Lemma 1 we have

$$\sum_{i=0}^{d} \underline{h}_i = \sum_{i=0}^{k} h_i + \sum_{i=k+1}^{d} \underline{m}^{\langle i/d\rangle} = \sum_{i=0}^{k} h_i + \sum_{i=0}^{d} \underline{m}^{\langle i/d\rangle} - \sum_{i=0}^{k} \underline{m}^{\langle i/d\rangle}$$

$$= \sum_{i=0}^{k} h_i + m^{\langle d/d\rangle} - m^{\langle k/d\rangle}$$

$$= f_d.$$

Now suppose we have $\underline{h}' = (h_0, \cdots, h_k, \underline{h}'_{k+1}, \cdots, h'_d)$ with $\underline{h}'_i = \underline{m}^{\langle i/d \rangle}$ for $i = k+1, \cdots, r-1$ and $\underline{h}'_r \leqq \underline{m}^{\langle r/d \rangle} - 1$. That is,

$$\underline{h}'_r \leqq \binom{m_d - 1 + r - d}{r} + \cdots + \binom{m_q - 1 + r - d}{q + r - d} - 1,$$

where $q = \max\{1 + d - r, l\}$. But then, for $j = r+1, \cdots, d$,

$$(\underline{h}'_r)^{\langle j/r \rangle} \leqq \binom{m_d - 1 + j - d}{j} + \cdots + \binom{m_q - 1 + j - d}{q + j - d} \leqq \underline{m}^{\langle j/d \rangle},$$

so that

$$\sum_{i=0}^{d} \underline{h}'_i \leqq \sum_{i=0}^{k} h_i + \sum_{i=k+1}^{r-1} \underline{m}^{\langle i/d \rangle} + \underline{m}^{\langle r/d \rangle} - 1 + \sum_{i=r+1}^{d} \underline{m}^{\langle i/d \rangle} \leqq \sum_{i=0}^{d} \underline{h}_i - 1 < f_d.$$

Thus, for $i = k+1, \cdots, d$, $\underline{h}_i = \underline{m}^{\langle i/d \rangle}$ is the minimum difference in Proposition 7.

The proof of Theorem 4 is now complete. □

REFERENCES

[1] A. AHO, J. E. HOPCROFT AND J. D. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[2] M. O. BALL, *The complexity of network reliability computations*, Networks, 10 (1980), pp. 153–165.

[3] ———, *Computing network reliability*, Oper. Res., 27 (1979), pp. 823–838.

[4] ———, *Network reliability analysis: Algorithms and complexity*, Ph.D. thesis, Cornell University, Ithaca, NY, 1977.

[5] M. O. BALL AND G. L. NEMHAUSER, *Matroids and a reliability analysis problem*, Math. Oper. Res., 4 (1979), pp. 132–143.

[6] M. O. BALL AND J. S. PROVAN, *Calculating bounds on reachability and connectedness in stochastic networks*, Working Paper MS/S 81–012, College of Business and Management, Univ. Maryland, College Park, March, 1981.

[7] N. BIGGS, *Algebraic Graph Theory*, Cambridge University Press, New York, 1974.

[8] A. BJORNER, *Some matroid inequalities*, Discrete Math., 31 (1980), pp. 101–103.

[9] Z. W. BIRNBAUM, J. D. ESARY AND S. C. SAUNDERS, *Multicomponent systems and structures and their reliability*, Technometrics, 3 (1961), pp. 55–71.

[10] R. BIXBY, *The minimum number of edges and vertices in a graph with edge connectivity N and M N-bonds*, Networks, 5 (1975), pp. 259–298.

[11] H. BRUGESSER AND P. MANI, *Shellable decompositions of cells and spheres*, Math. Scand., 29 (1971), pp. 197–205.

[12] J. A. BUZACOTT, *A recursive algorithm for finding the probability that a graph is disconnected*, Networks, 10 (1980), pp. 311–328.

[13] G. DANARAJ AND V. KLEE, *Shellings of spheres and polytopes*, Duke Math. J., 41 (1974), pp. 443–451.

[14] J. D. ESARY AND F. PROSCHAN, *A reliability bound for systems of maintained independent components*, J. Amer. Statist. Assoc., 65 (1970), pp. 329–338.

[15] S. EVEN AND R. E.TARJAN, *Network flow and testing graph connectivity*, SIAM J. Comput., 4 (1975), pp. 507–518.

[16] G. KATONA, *A theorem of finite sets*, in Theory of Graphs, Proc. Tihany Colloquium, Sept. 1966, P. Erdos and G. Katona, eds., 1966, pp. 209–214.

[17] J. B. KRUSKAL, *The number of simplices in a complex*, in Mathematical Optimization Techniques, R. Bellman, ed., Univ. California Press, Berkeley, 1963, pp. 251–278.

[18] P. MCMULLEN, *Convex Polytopes and the Upper Bound Conjecture*, Cambridge University Press, New York, 1971.

[19] M. MESSINGER AND M. SHOOMAN, *Reliability approximations for complex structures*, IEEE Proc. Symposium on Reliability, Washington, DC, 1967, pp. 292–301.

[20] J. S. PROVAN AND M. O. BALL, *The complexity of counting cuts and computing the probability that a graph is connected*, Working Paper MS/S 81-002, College of Business and Management, University of Maryland, College Park, January, 1981.

[21] J. S. PROVAN AND L. J. BILLERA, *Decompositions of simplicial complexes related to diameters of convex polyhedra*, Math. Oper. Res., 5 (1980), pp. 579–594.

[22] ———, *Simplicial complexes associated with convex polyhedra* I: *Constructions and examples*, Tech. Rep. No. 402, School of Operations Research and Industrial Engineering, Cornell Univ., Ithaca, NY, 1979.

[23] J. RIORDAN, *Combinatorial Identities*, John Wiley, New York, 1968.

[24] A. ROSENTHAL, *A computer scientist looks at reliability and fault tree analysis*, in Reliability and Fault Tree Analysis, R. E. Barlow, J. B. Fussell and N. D. Singpurwalla, eds. Society for Industrial and Applied Mathematics, Philadelphia, 1975, pp. 133–152.

[25] ———, *Computing the reliability of complex networks*, SIAM J. Appl. Math., 32 (1977), pp. 384–393.

[26] R. P. STANLEY, *Balanced Cohen-Macauley complexes*, Trans. Amer. Math. Soc., 249 (1979), pp. 138–151.

[27] ———, *Cohen-Macauley complexes*, in Higher Combinatorics, M. Aigner, ed., D. Reidel, Dordrecht, Holland, (1977), pp. 51–62.

[28] L. G. VALIANT, *The complexity of enumeration and reliability problems*, SIAM J. Comput., 8 (1979), pp. 410–421.

[29] R. M. VAN SLYKE AND H. FRANK, *Network reliability analysis* 1, Networks, 1 (1972), pp. 279–290.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 2, June 1982

0196-5212/82/0302-0004 $01.00/0

AN EFFECTIVE FORMULA FOR THE NUMBER OF SOLUTIONS OF LINEAR BOOLEAN EQUATIONS*

P. L. BUZYTSKY†

Abstract. The paper is devoted to establishing an approximate formula for the number of solutions of one boolean linear equation. The formula is deduced by using a technique of analytical number theory.

Let us consider the equation

$$a_1x_1+\cdots+a_nx_n=b, \tag{1}$$

where a_j, b are positive integers, $x_j=0, 1$, $j=1,\cdots,n$ and $b\leqq\frac{1}{2}\sum_{j=1}^{n}a_j$. The analytical approach to analysis of (1) was first suggested in [1].

This paper is the next step in the course of studying integer programming problems by means of analytical methods. The research work was suggested by G. Freiman, and is being conducted under his guidance (see [2]–[5]). In [2] there is an asymptotic formula for the number of solutions of one Boolean equation (1), and the other papers are devoted to the analysis of this formula and its applicability.

The aim of this paper is to establish a new formula which should be convenient for practical calculations.

The number of solutions of (1) I_n is expressed as

$$I_n=\exp(\sigma b)\prod_{j=1}^{n}(1+\exp(-\sigma a_j))\int_0^1\prod_{j=1}^{n}(p_{1j}+p_{2j}\exp(2\pi i\alpha a_j))\exp(-2\pi i\alpha b)\,d\alpha, \tag{2}$$

where

$$p_{1j}=\frac{1}{1+\exp(-\sigma a_j)},\qquad p_{2j}=\frac{\exp(-\sigma a_j)}{1+\exp(-\sigma a_j)}.$$

The formula (2) is true for any real σ. We determine σ as a solution of the equation

$$\sum_{j=1}^{n}\frac{a_j}{(\exp(\sigma a_j)+1)}=b. \tag{3}$$

It is not difficult to see that (3) has a unique solution.

We introduce the following notation (see also [2]):

$$D=\sum_{j=1}^{n}p_{1j}p_{2j}a_j^2,$$

$$h=\max_j p_{2j}\frac{a_j^2}{D},$$

$$\rho_3=\sum_{j=1}^{n}a_j^3p_{2j},$$

$$t=\frac{1}{cD^{1/2}},\qquad c \text{ a positive constant},$$

$$\nu=\sqrt{2\pi D}\int_t^{1/2}\left|\prod_{j=1}^{n}(p_{1j}+p_{2j}\exp(2\pi i\alpha a_j))\right|d\alpha,$$

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x\exp\left(-\frac{z^2}{2}\right)dz.$$

* Received by the editors November 8, 1980, and in final form June 24, 1981.

† Academy of Sciences of the USSR, Central Institute of Economics and Mathematics, Vavilov Street 44-2. Moscow V333 USSR.

The integral that appears in (2) can be partitioned as follows:

$$I=\int_{-1/2}^{1/2}\prod_{j=1}^{n}(p_{1j}+p_{2j}\exp(2\pi i\alpha a_j))\exp(-2\pi i\alpha b)\,d\alpha=\int_{-t}^{t}+\left(\int_{-1/2}^{-t}+\int_{t}^{1/2}\right)=I_1+I_2.$$

Transforming the integral I_1 taking into account (3) and denoting

$$\mu_j=2\pi\alpha a_j,$$

we obtain

$$\begin{aligned}
I_1&=\int_{-t}^{t}\prod_{j=1}^{n}(p_{1j}+p_{2j}\exp(i\mu_j))\prod_{j=1}^{n}\exp(-i\mu_j p_{2j})\,d\alpha\\
&=\int_{-t}^{t}\prod_{j=1}^{n}(p_{1j}\exp(-i\mu_j p_{2j})+p_{2j}\exp(i\mu_j p_{1j}))\,d\alpha\\
&=\int_{-t}^{t}\prod_{j=1}^{n}\exp(-\tfrac{1}{2}\mu_j^2 p_{1j}p_{2j})\prod_{j=1}^{n}\frac{p_{1j}\exp(-i\mu_j p_{2j})+p_{2j}\exp(i\mu_j p_{1j})}{\exp(-\frac{1}{2}\mu_j^2 p_{1j}p_{2j})}\,d\alpha\\
&=\int_{-t}^{t}\exp\left(-\tfrac{1}{2}\sum_{j=1}^{n}\mu_j^2 p_{1j}p_{2j}\right)d\alpha+R=\int_{-t}^{t}\exp(-2\pi^2\alpha^2 D)\,d\alpha+R,
\end{aligned}\tag{4}$$

where

$$R=\int_{-t}^{t}\exp(-2\pi^2\alpha^2 D)\left(\prod_{j=1}^{n}\frac{p_{1j}\exp(-i\mu_j p_{2j})+p_{2j}\exp(i\mu_j p_{1j})}{\exp(-\frac{1}{2}\mu_j^2 p_{1j}p_{2j})}-1\right)d\alpha.$$

Denoting

$$\Delta_j=\frac{p_{1j}\exp(-i\mu_j p_{2j})+p_{2j}\exp(i\mu_j p_{1j})}{\exp(-\frac{1}{2}\mu_j^2 p_{1j}p_{2j})}-1,$$

we find

$$\begin{aligned}
\left|\prod_{j=1}^{n}(1+\Delta_j)-1\right|&=|\Delta_1+\cdots+\Delta_n+\Delta_1\Delta_2+\cdots|\\
&\leqq|\Delta_1|+\cdots+|\Delta_n|+|\Delta_1\Delta_2|+\cdots\\
&=\left(\prod_{j=1}^{n}(1+|\Delta_j|)-1\right)\leqq\exp\left(\sum_{j=1}^{n}|\Delta_j|\right)-1,
\end{aligned}$$

so that we have

$$|R|\leqq\int_{-t}^{t}\exp(-2\pi^2\alpha^2 D)\left(\exp\left(\sum_{j=1}^{n}|\Delta_j|\right)-1\right)d\alpha.\tag{5}$$

We now bound the value $|\Delta_j|$ from above

$$\begin{aligned}
|\Delta_j|&=\left|\frac{p_{1j}\exp(-i\mu_j p_{2j})+p_{2j}\exp(i\mu_j p_{1j})-\exp(-\frac{1}{2}\mu_j^2 p_{1j}p_{2j})}{\exp(-\frac{1}{2}\mu_j^2 p_{1j}p_{2j})}\right|\\
&\leqq\frac{|A_j(\mu_j)-E_j(\mu_j)|+|B_j(\mu_j)|}{E_j(\mu_j)},
\end{aligned}\tag{6}$$

where

$$A_j(\mu_j) = p_{1j}\cos(\mu_j p_{2j}) + p_{2j}\cos(\mu_j p_{1j}),$$
$$B_j(\mu_j) = p_{1j}\sin(\mu_j p_{2j}) - p_{2j}\sin(\mu_j p_{1j}),$$
$$E_j(\mu_j) = \exp(-\tfrac{1}{2}\mu_j^2 p_{1j}p_{2j}).$$

Using Taylor's formula, we have, for $|B_j(\mu_j)|$,

$$|B_j(\mu_j)| = \left| B_j(0) + B_j'(0)\mu_j + \frac{B_j''(0)\mu_j^2}{2!} + \frac{B_j'''(\xi)\mu_j^3}{3!} \right|,$$

where $0 \leqq \xi \leqq \mu_j$.

Since $B_j(0) = B_j'(0) = B_j''(0) = 0$, we obtain

$$|B_j(\mu_j)| \leqq \frac{|\mu_j|^3 p_{1j}p_{2j}}{3!} \leqq \frac{|\mu_j|^3 p_{2j}}{6} = \gamma_j. \tag{7}$$

We now bound the value $|f_j(\mu_j)| = |A_j(\mu_j) - E_j(\mu_j)|$. Since

$$f_j(0) = f_j'(0) = f_j''(0) = 0,$$

then

$$f_j(\mu_j) = \frac{|\mu_j|^3}{3!}\left| p_{1j}p_{2j}(p_{2j}^2\sin(\xi p_{2j}) + p_{1j}^2\sin(\xi p_{1j})) \right.$$
$$\left. - \xi p_{1j}^2 p_{2j}^2 \exp(-\tfrac{1}{2}\xi^2 p_{1j}p_{2j})(3 - \xi^2 p_{1j}p_{2j}) \right|,$$

where $0 \leqq \xi \leqq \mu_j$. Since $p_{1j}p_{2j} < 1$ and $y(3-y^2)/\exp(y^2/2) < 2$ for any y, we have

$$|f_j(\mu_j)| \leqq \frac{|\mu_j|^3}{3!}(p_{1j}p_{2j}(p_{1j}^2 + p_{2j}^2 + 2p_{1j}p_{2j})) = \gamma_j. \tag{8}$$

From, (6), (7) and (8), we obtain

$$|\Delta_j| < \frac{2\gamma}{E_j}. \tag{9}$$

We return now to the bound (5) for $|R|$.

$$\begin{aligned} |R| &\leqq \int_{-t}^{t} \exp(-2\pi^2\alpha^2 D)\left(\exp\left(\textstyle\sum_{j=1}^n |\Delta_j|\right) - 1\right) d\alpha \\ &= \frac{1}{\pi D^{1/2}} \int_0^{2\pi/c} \exp(-\tfrac{1}{2}\beta^2)\left(\exp\left(\textstyle\sum_{j=1}^n |\Delta_j|\right) - 1\right) d\beta. \end{aligned} \tag{10}$$

By virtue of inequalities (7), (8) and (9), we obtain

$$|\Delta_j| < \exp\left(\frac{\frac{1}{2}\beta^2 a_j^2 p_{1j}p_{2j}}{D}\right)\frac{\frac{1}{3}\beta^3 a_j^3 p_{2j}}{D^{3/2}},$$

and

$$\sum_{j=1}^{n} |\Delta_j| \leqq \exp\left(\frac{\beta^2 h}{2}\right)\frac{\frac{1}{3}\beta^3 \rho_3}{D^{3/2}}. \tag{11}$$

With

$$\Delta_0 = \exp\left(\frac{2\pi^2 h}{c^2}\right)\frac{8\pi^3\rho_3}{3c^3D^{3/2}},$$

and

$$q = \frac{\exp(\Delta_0)-1}{\Delta_0}$$

using

$$z \leqq z_0 \Rightarrow \exp(z)-1 \leqq z\,\frac{\exp(z_0)-1}{z_0}$$

we obtain, from (10) and (11),

$$\begin{aligned}|R| &\leqq \frac{q}{\pi D^{1/2}}\int_0^{2\pi/c} \exp\left(-\tfrac{1}{2}\beta^2\right)\exp\left(\tfrac{1}{2}\beta^2 h\right)\frac{\rho_3}{3D^{3/2}}\beta^3\,d\beta \\ &= \frac{2q\rho_3}{3\pi D^{1/2}(1-h)^2D^{3/2}}\left(1-\exp\left(\frac{-2\pi^2(1-h)}{c^2}\right)\right)\left(1-\frac{2\pi^2(1-h)}{c^2}\right)\right) \qquad (12)\\ &= u/\sqrt{2\pi D}.\end{aligned}$$

Further, we have

$$\int_t^\infty \exp\left(-2\pi^2\alpha^2 D\right)d\alpha = \frac{\frac{1}{2}-\Phi(2\pi/c)}{\sqrt{2\pi D}}. \qquad (13)$$

Therefore, (4), (12) and (13) imply

$$I_1 = \frac{1}{\sqrt{2\pi D}}\left(1+2\theta\left(\left(\frac{1}{2}-\Phi\left(\frac{2\pi}{c}\right)\right)+\frac{u}{2}\right)\right),$$

where $|\theta|<1$, and ultimately,

$$I_n = \exp(\sigma b)\prod_{j=1}^{n}\left(1+\exp(-\sigma a_j)\right)\frac{1}{\sqrt{2\pi D}}\left(1+2\theta\left(\left(\frac{1}{2}-\Phi\left(\frac{2\pi}{c}\right)\right)+\frac{u}{2}+\nu\right)\right). \qquad (14)$$

Some words should be said about numerical usability of the formula (14). It is easy to see that the required σ can be easily found from (3) by means of any appropriate technique; for instance, one may use bisection. Then there is no problem in computing D, and thus the main term of the formula, since we need to compute only the product of n terms in it. As far as the remainder term is concerned, there are three terms in it. $\Phi(x)$ is a tabulated function with known values. Given σ, the computation of u requires $O(n)$ operations, for we need only D, h and ρ_3. The only difficulty is with term ν. In [3], there is an approach to computing ν by means of a ε-net with an estimated efficiency. However, as pointed out above, the share of (1) with large ν can be neglected (see [4]). So, for practical purposes, we may assume that ν is sufficiently small. Therefore, formula (14) can be readily used in practical problems.

REFERENCES

[1] A. A. BERNSTEIN, *A numerical method of solving an allocation problem*, Coll. Problems of Program-Goal Planning and Control, CEMI, Moscow, 1978. (In Russian).

[2] G. A. FREIMAN, *An analytical method of analysis of linear boolean equations*, Ann. New York Academy of Sciences, 337, 1980.
[3] P. L. BUZYTSKY AND G. A. FREIMAN *On application of analytical methods in combinatorial problems*, Izv. Akad. Nauk SSSR, Tekhnicheskaya Kibernetika, N2 1980, (In Russian).
[4] ———, *On the possibilities of solving combinatorial problems by analytic methods*, Ann. New York Academy of Sciences, 337, 1980.
[5] ———, *Analytical methods in integer programming*, Preprint, CEMI, Moscow, 1980. (In Russian).

SIAM J. ALG. DISC. METH.
Vol. 3, No. 2, June 1982

0196-5212/82/0302-0005 $01.00/0

FAREY SERIES AND MAXIMAL OUTERPLANAR GRAPHS*

CHARLES J. COLBOURN†

Abstract. Certain graphs representing Farey series of irreducible fractions are shown to be maximal outerplanar. For a suitable generalization of Farey series, the class of graphs obtained is exactly the class of maximal outerplanar graphs. Using a representation of maximal outerplanar graphs as series of irreducible fractions, efficient algorithms for deciding isomorphism of maximal outerplanar graphs and for deciding whether one maximal outerplanar graph is a subgraph of another are described.

1. Introduction. Recently, Matula and Kornerup [6] introduced Farey fraction graphs, a graph representation of the Farey series from classical number theory. They demonstrated that Farey fraction graphs are uniquely and minimally 3-colorable, uniquely Hamiltonian and perfect.

In this paper we show that Farey fraction graphs are, in fact, maximal outerplanar graphs; from this observation, Matula and Kornerup's results follow immediately. We then introduce a natural generalization which we call Farey graphs and demonstrate that the class of Farey graphs is exactly the class of maximal outerplanar graphs. The proof that all maximal outerplanar graphs are Farey graphs yields a remarkable canonical form for maximal outerplanar graphs. Applications of this canonical form to deciding isomorphism of maximal outerplanar graphs and to deciding whether one maximal outerplanar graph is a subgraph of another in $O(n^2)$ time are given.

2. Definitions. Number-theoretic definitions can be found in [5], graph-theoretic ones in [2], [4]. The *Farey series* F_n is the series of all irreducible fractions between 0/1 and 1/1 with denominator not exceeding n. For example, $F_5 = 0/1, 1/5, 1/4, 1/3, 2/5, 1/2, 3/5, 2/3, 3/4, 4/5, 1/1$. A *punctured Farey series* is a series obtained from a Farey series by any number of the following *puncturing* operations: select three fractions h'/k', h/k, h''/k'' which are adjacent in the series and for which $k > k'$ and $k > k''$; then delete h/k from the series. For example, 0/1, 1/3, 1/2, 1/1 is a punctured Farey series, whereas 0/1, 1/3, 2/3, 1/1 is not. A *Farey graph* is a graph whose vertices are the fractions in a punctured Farey series. Two vertices h/k and h'/k' are adjacent if and only if $|hk' - h'k| = 1$.

An *outerplanar* graph is a graph which can be embedded in the plane so that no two edges cross ("planar") and every vertex lies on the exterior face ("outer"). A *maximal outerplanar* graph is an outerplanar graph to which no edge can be added without destroying outerplanarity.

A result of Tang [9] has as a corollary that maximal outerplanar graphs are uniquely Hamiltonian. A result of Read [7] demonstrates that the chromatic polynomial of any n-vertex maximal outerplanar graph M is $P(M, \lambda) = \lambda(\lambda - 1)(\lambda - 2)^{n-2}$, and hence they are 3-chromatic. It is straightforward to verify that maximal outerplanar graphs are perfect and uniquely 3-colorable.

3. Farey graphs and maximal outerplanar graphs. In this section we prove that the class of Farey graphs is precisely the class of maximal outerplanar graphs.

THEOREM 1. *Farey graphs are maximal outerplanar. Further, the unique Hamilton cycle involves all edges connecting pairs of neighboring elements in the punctured Farey series.*

* Received by the editors August 28, 1979, and in final form August 17, 1981.

† Department of Computational Science, University of Saskatchewan, Saskatoon, Saskatchewan, S7N OWO, Canada.

Proof. Suppose we are given a punctured Farey series F and its Farey graph $G(F)$. Let F' be obtained from F by one puncture; let $G(F')$ be its Farey graph.

We will prove the theorem by induction on the number of terms in the punctured Farey series. If $|F| \leqq 3$, the result is trivial. We therefore suppose that all Farey graphs with $|F|-1$ vertices are maximal outerplanar. By induction, then, $G(F')$ is maximal outerplanar. Let h/k be the element in $F-F'$; let h'/k' and h''/k'' be its neighboring elements in the series F. Elementary number-theoretic considerations [5] demonstrate that h/k, h'/k' and h''/k'' induce a triangle in $G(F)$. We will next prove that h/k has no adjacencies in $G(F)$ other than h'/k' and h''/k''. It follows from [5, Thm. 29] that $h = h' + h''$ and $k = k' + k''$. Suppose the vertex y/z is connected to h/k. Then $|(h'+h'')z - (k'+k'')y| = 1$, so $|h'z - k'y| + |h''z - k''y| = 1$. Since $|h'z - k'y| = 0$ only if $h'/k' = y/z$, we conclude that y/z is one of h'/k' or h''/k''.

By induction, the edge connecting h'/k' and h''/k'' is on the unique Hamilton cycle of $G(F')$. Hence, $G(F)$ is maximal outerplanar and $(h'/k', h/k)$ and $(h/k, h''/k'')$ lie on the unique Hamilton cycle of $G(F)$. □

THEOREM 2. *Maximal outerplanar graphs are Farey graphs.*

Proof. Given a maximal outerplanar graph M, we show how to find a punctured Farey series whose Farey graph is M. We do this by labeling the vertices of M with irreducible fractions, in the following manner.

Select an arbitrary exterior edge and label its endpoints $0/1$ and $1/1$. Let all edges be untagged. Now until all vertices are labeled, select an untagged edge with both endpoints v and w labeled h/k and h'/k'. If there is no unlabeled vertex adjacent to both v and w, tag the edge and return to select another. Otherwise, observe that there is a unique unlabeled vertex z adjacent to both v and w. We label z with $h+h'/k+k'$, tag the edge and return to select another.

It is immediate [5] that the fractions assigned form a punctured Farey series; this series depends only on the selection of vertices to be labeled $0/1$ and $1/1$. □

Theorems 1 and 2, together with the current knowledge concerning maximal outerplanar graphs, supply simpler proofs of Matula and Kornerup's observations about Farey graphs.

4. Applications to isomorphism. A labeling assigned by the method in the proof of Theorem 2 is called a *Farey labeling*; it is interesting to note that knowing just the set of Farey labels (called the graph's Farey form) determines the graph uniquely. An *n-vertex* maximal outerplanar graph can have potentially $O(n)$ different Farey forms.

A canonical representation for a maximal outerplanar graph can be found by finding its lexicographically smallest Farey form; this is the graph's *Farey code.* Two maximal outerplanar graphs are isomorphic if and only if they have the same Farey code. This gives an $O(n^2)$ algorithm for deciding isomorphism—$O(n)$ ways of selecting vertices to be labeled $0/1$ and $1/1$ and, after this choice, $O(n)$ time to compute the remainder of the Farey labeling. This method is not as efficient as the known algorithms for deciding isomorphism of maximal outerplanar graphs [1], [3], [8], but is interesting nonetheless in that it is a fundamentally different approach.

The primary advantage of the Farey scheme is that it generalizes immediately to the problem of determining whether one maximal outerplanar graph is a subgraph of another. We will elaborate on this here. We define a *portion* of a maximal outerplanar graph to be the result of selecting two vertices v and w connected by an interior edge, then deleting all vertices strictly between v and w on one of the two sections of the unique Hamilton cycle from v to w. The vertices v and w are labeled "start" and "finish", respectively.

A maximal outerplanar graph with n vertices has only $O(n)$ interior edges; thus, it has only $O(n)$ portions.

THEOREM 3. *A maximal outerplanar graph M is a subgraph of a maximal outerplanar graph N if and only if either*

(1) *the Farey code of M is a subset of some Farey form of N,* or

(2) *the Farey code of M is a subset of the Farey form of some portion of N in which "start" is given label $0/1$ and "finish" label $1/1$.*

Proof. The conditions are clearly sufficient: we will show that they are also necessary. Let the vertices v and w be the vertices labeled $0/1$ and $1/1$ in the Farey code of M. Suppose M is a subgraph of N, and let f: $V(M) \to V(N)$ be an embedding of M into N. If $(f(v), f(w))$ is an exterior edge of N, there is a Farey form of N in which $f(v)$ is labeled $0/1$ and $f(w)$ is labeled $1/1$. This Farey form contains the Farey code of M, and thus, condition (1) is satisfied. Otherwise, $(f(v), f(w))$ is an interior edge of N. In this case, M is a subgraph of one of the two portions of N determined by the edge $(f(v), f(w))$. Then there is a Farey form of one of these portions having $f(v)$ labeled $0/1$ and $f(w)$ labeled $1/1$ which has the Farey code of M as a subset; hence, condition (2) is satisfied. □

COROLLARY 4. *There is an $O(n^2)$ algorithm to decide whether a maximal outerplanar graph is a subgraph of an n-vertex maximal outerplanar graph.*

Proof. Our earlier remarks show that condition (1) can be checked in $O(n^2)$ time. Further, since there are only $O(n)$ portions and the Farey labeling for each can be completed in $O(n)$ time, condition (2) can also be checked in $O(n^2)$ time. □

One remark is in order. The Farey labels assigned in computing Farey forms might have as many as $O(n)$ bits, thus, requiring $O(n^2)$ bit operations to examine a Farey form. In the complexity statements in this work we ignore this issue and assume that our model of computation can manipulate integers with $O(n)$ bits in a single step.

5. Conclusions. Perhaps the most interesting aspect of this work is that Farey series, a natural classical concept in number theory, correspond to maximal outerplanar graphs, a natural concept in graph theory. We expect that graph-theoretic investigations of other number-theoretic concepts will reveal many such natural correspondences.

REFERENCES

[1] T. BEYER, W. JONES AND S. L. MITCHELL, *A linear algorithm for isomorphism of maximal outerplanar graphs*, Tech. Report CS-TR-78-01, Univ. of Oregon, Eugene, 1978.

[2] J. A. BONDY AND U.S.R. MURTY, *Graph Theory with Applications*, Macmillan, London, 1976.

[3] C. J. COLBOURN AND K. S. BOOTH, *Linear automorphism algorithms for trees, interval graphs and planar graphs*, SIAM J. Comput. 10 (1981), pp. 203–225.

[4] F. HARARY, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.

[5] G. H. HARDY AND E. M. WRIGHT, *An Introduction to the Theory of Numbers*, Oxford University Press, Oxford, 1962.

[6] D. W. MATULA AND P. KORNERUP, *A graph theoretic interpretation of fractions, continued fractions and the GCD algorithm*, Proc. Tenth Southeastern Conference on Combinatorics, Graph Theory, and Computing, 1979, p. 932.

[7] R. C. READ. *An introduction to chromatic polynomials*, J. Combin. Theory, 4 (1968), pp. 52–71.

[8] M. M. SYSŁO, *Linear time algorithm for coding outerplanar graphs*, Report N-20, Institute of Computer Science, Wrocław University, Wrocław, Poland, 1977.

[9] D. T. TANG, *Bi-path networks and multicommodity flows*, IEEE Trans. Circuit Theory, CT-11 (1964), pp. 468–474.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 2, June 1982

0196-5212/82/0302-0006 $01.00/0

SCHEDULING TO MAXIMIZE THE MINIMUM PROCESSOR FINISH TIME IN A MULTIPROCESSOR SYSTEM*

BRYAN L. DEUERMEYER,† DONALD K. FRIESEN‡ AND MICHAEL A. LANGSTON§

Abstract. This investigation considers the problem of nonpreemptively assigning a set of independent tasks to a system of identical processors to maximize the earliest processor finishing time. While this goal is a nonstandard scheduling criterion, it does have natural applications in certain maintenance scheduling and deterministic fleet sizing problems. The problem is NP-hard, justifying an analysis of heuristics such as the well-known LPT algorithm in an effort to guarantee near-optimal results. It is proved that the worst-case performance of the LPT algorithm has an asymptotically tight bound of $\frac{4}{3}$ times the optimal.

1. Introduction. Consider M identical processors and a set of N independent tasks $T=\{t_1, \cdots, t_N\}$ each having a processing time given by $l(t_i)$, where $l: T \to \mathbb{R}^+$. The central problem we address is the nonpreemptive scheduling of the tasks in an effort to fully occupy each processor for as long as possible. This objective is quite distinct from the usual performance measure of minimizing the makespan (final task completion time). Herein we concentrate on keeping the processors busy by maximizing the minimum processor completion time α_{ALG}, where the finish time of the jth processor is the time at which the last task to be executed by processor j is completed.

This problem was initially motivated by investigations into the sequencing of maintenance actions for modular gas turbine aircraft engines. In the simplest example of this problem, suppose a fleet of M identical machines (engines) must be kept operational for as long as possible, and each machine requires the same life-limited part. Further suppose that N spares (with potentially different field-lives) of this part are initially available. Then the problem of sequencing the replacements to maximize the total time the fleet is operational is identical to the scheduling problem stated above. This maintenance problem has been studied by two of the authors in [3]. It is important to note here that this maintenance scheduling problem is unlike other maintenance problems currently found in the literature—which in part leads to the unusual performance criterion employed.

The scheduling problem stated above bears similarity to at least two common scheduling problems previously studied in the literature. The objective of our problem is somewhat related to minimizing the maximum earliness, a "nonregular" measure of performance identified in [2], but the absence of task deadlines changes the problem substantially. The fleet sizing problem [6] shares our motivation but differs in that its representation as a scheduling problem would necessitate deterministic demands, unit-time tasks and a uniform processor system (i.e., processors of different speeds).

Maximizing the minimum processor finish time, like a host of other scheduling and related combinatorial problems (in particular, the makespan problem), is NP-hard. To see this it is necessary to reduce, in polynomial time, some other problem known to be NP-complete to a version of this problem. Specifically, it is not difficult to reduce the partition problem to a "yes–no" version of this problem on two processors. It is generally considered unlikely that an NP-complete or harder problem will permit an efficient (i.e., polynomial-time) solution procedure. Therefore, it is common to solve

* Received by the editors April 21, 1981.

† Department of Industrial Engineering, Division of Industrial Engineering, Texas A&M University, College Station, Texas 77843.

‡ Department of Industrial Engineering, Division of Computing Science, Texas A&M University, College Station, Texas 77843.

§ Department of Computer Science, Washington State University, Pullman, Washington 99164.

these problems by efficient heuristic methods in hopes of finding "good" solutions rather than attempting to find optimal solutions. In this paper we analyze the LPT (longest processing time first) algorithm, since it is known to work well on the makespan problem.

An important aspect in the analysis of a heuristic procedure is to determine the quality of its performance relative to optimal solutions. Worst-case analysis yields a relatively simple measure of performance of a heuristic algorithm. The focus of this article is to characterize the worst-case behavior of the LPT algorithm applied to the problem introduced above. The LPT procedure sorts the task set T into a nonincreasing sequence and then serially assigns each task to the next available processor with ties broken arbitrarily. For the makespan problem on M processors, Graham [5] proved that LPT guarantees a tight worst-case performance of $(\frac{4}{3}-1/3M)$ times the optimal value. We expected LPT to perform in a similar way on our problem, and in fact, we prove that $\frac{4}{3}$ is an asymptotically tight worst-case bound. However, our analysis is quite unlike Graham's. In particular, we must explicitly deal with tasks having "small" processing times, a consideration easily avoided in his proof.

At first glance it may appear that our problem is a "dual" to the makespan problem, since one would expect heuristics to balance the work load over all processors; the objectives differ in that one is a maximization and the other is a minimization problem. It seems reasonable to suggest that good algorithms for the makespan problem should produce good solutions to the problem considered in this paper. However, there are important differences in the worst-case behavior of algorithms applied to these two problems. Consider, for example, the MULTIFIT algorithm of Coffman, Garey and Johnson [1]. In the solution of the makespan problem using MULTIFIT, it is easy to construct examples where one processor is never used. Such a solution is tolerable for the makespan problem but is totally unacceptable for our problem. Modifications of MULTIFIT can be devised which would be more suitable for our problem, but we could find none which produces a better worst-case bound than that of LPT. Consequently we have limited our analysis to a proof of the worst-case bound for the LPT algorithm.

In what follows, an instance I refers to a particular choice of problem parameters, $I=(N, M, T, l)$. Given any I, we use P_i, $1 \leqq i \leqq M$, to represent both the ith processor and its subschedule (those tasks assigned to processor i) in the LPT algorithm. Similarly, P_j^*, $1 \leqq j \leqq M$, denotes an optimal subschedule for processor j. For any $T' \subseteq T$, the length of T' is given by

$$l(T')=\sum_{x \in T'} l(x).$$

For a given problem instance, I, $\alpha_{\mathrm{LPT}}(I)$ and $\alpha_{\mathrm{OPT}}(I)$ denote the minimum processor finish time of the LPT and optimal schedules, respectively. The worst-case bound $Q_M(\mathrm{LTP})$ for the LPT algorithm on M processors is defined by

$$Q_M(\mathrm{LPT})=\sup\{\alpha_{\mathrm{OPT}}(I)/\alpha_{\mathrm{LPT}}(I);\ I \text{ has } M \text{ processors}\}.$$

The remainder of this paper is organized as follows. The next section shows by means of an example that $\frac{4}{3}$ is an asymptotic lower bound on $Q_M(\mathrm{LTP})$. In § 3 we assume the existence of a counterexample to the claim that $\frac{4}{3}$ is an upper bound on $Q_M(\mathrm{LTP})$ and, hence, the existence of a "minimal" counterexample whose properties we analyze. Section 4 contains our main result—we establish a contradiction to the presumed existence of a counterexample and, thus, prove that $\frac{4}{3}$ is a "tight" bound for the worst-case performance of the LPT algorithm.

2. Lower bound. In this section we demonstrate that the worst-case bound derived in § 4 is asymptotically tight. Example 2.1 below depicts a family of problem instances for which $\alpha_{\mathrm{OPT}}(I)/\alpha_{\mathrm{LPT}}(I)$ exceeds any real number smaller than $\frac{4}{3}$. Figure 2.1 (with time the vertical axis and processors the horizontal axis) illustrates an LPT schedule and an optimal schedule in the case where M is even. A slightly different picture is required when M is odd.

Example 2.1. Define an instance I by choosing $M \geqq 2$, $N = 3M-1$, $l(t_i) = M$, $i > 2M$ and $l(t_i) = 2M - \mathrm{FLOOR}\,((i+1)/2)$, $i = 1, 2, \cdots, 2M$. Then (see Fig. 2.1 for even M)

$$Q_M(\mathrm{LPT}) \geqq \alpha_{\mathrm{OPT}}(I)/\alpha_{\mathrm{LPT}}(I) = (4M-2)/(3M-1) = \frac{4}{3} - \frac{2}{3(3M-1)}.$$

M	M	M	M	$\cdots$	M	
M	M	$M+1$	$M+1$	$\cdots$	$\frac{3M}{2}-1$	$\frac{3M}{2}-1$
$2M-1$	$2M-1$	$2M-2$	$2M-2$		$\frac{3M}{2}$	$\frac{3M}{2}$

(a)
LPT *schedule*: $\alpha_{\mathrm{LPT}}(I) = 3M-1$

$2M-1$	M	M	M	$\cdots$	M	M
	M	M	$M+1$	$\cdots$	$\frac{3M}{2}-2$	$\frac{3M}{2}-1$
$2M-1$	$2M-2$	$2M-2$	$2M-3$		$\frac{3M}{2}$	$\frac{3M}{2}-1$

(b)
optimal schedule: $\alpha_{\mathrm{OPT}}(I) = 4M-2$

FIG. 2.1. *Graphical representation of Example* 2.1 *for even M.*

3. Description of a minimal counterexample. We now assume the existence of a counterexample to our claim that $Q_M(\mathrm{LTP}) \leqq \frac{4}{3}$. Section 4 provides a proof that such a counterexample cannot exist. Specifically, we assume the existence of an instance $I^c = (N, M, T^c, l)$ such that

(i) $\alpha_{\mathrm{OPT}}(I^c)/\alpha_{\mathrm{LTP}}(I^c) > \frac{4}{3}$;

(ii) the instance is minimal in the sense that no fewer than M processors can be used to generate a counterexample;

(iii) the instance is minimal in the sense that no fewer than N tasks can be used to generate a counterexample on M processors.

Without loss of generality, we normalize the task lengths so that for an arbitrary optimal subschedule $l(P_j^*) \geqq 4$, $1 \leqq j \leqq M$, and $l(P_i) < 3$ for at least one LPT subschedule, P_i. We remark at this point that a simple "conservation of task length"

argument shows that there must be an LPT subschedule P_k for this normalized instance with $l(P_k) > 4$.

DEFINITION 3.1. A subschedule P_i is *dominated by* a subschedule P_j (P_i and P_j are typically generated by different algorithms) if there is a function f which maps tasks assigned to P_i into disjoint subsets of the tasks in P_j such that for any $x \in P_i$, $x = f(x)$ if $x \in P_j$ and $l(x) \leqq l(f(x))$ otherwise.

The rest of this section will present a number of properties the LPT and optimal schedules must have for the counterexample I^c.

LEMMA 3.1. *Let P_i be any* LPT *subschedule on I^c with $l(P_i) \geqq 3$. Then no subschedule of an optimal schedule on I^c will dominate P_i.*

Proof. Suppose the lemma is false and there is an optimal subschedule P_j^* which dominates P_i. Define a new instance of the problem $I^{c'} = (N', M', T^{c'}, l)$, where $M' = M - 1$, $T^{c'} = T^c - P_i$ and $N' = |T^{c'}|$. We obtain a schedule for $I^{c'}$ from the optimal schedule for I^c by filling each position formerly occupied by a task from P_i with that task's image under f and reassigning any remaining elements of P_j^* arbitrarily. The instance $I^{c'}$ uses fewer processors than I^c, $\alpha_{\mathrm{LPT}}(I^{c'}) = \alpha_{\mathrm{LPT}}(I^c)$ and $\alpha_{\mathrm{OPT}}(I^{c'}) \geqq \alpha_{\mathrm{OPT}}(I^c)$, contradicting the presumed minimality of I^c. □

LEMMA 3.2. *Every task t of T^c must have $l(t) < 3$.*

Proof. Suppose on the contrary there is a task t in T^c such that $l(t) \geqq 3$. Let P_i and P_j^* denote the subschedules containing task t under the LPT and optimal schedules, respectively. Since LPT can assign no additional tasks to any subschedule once its length is at least 3 (otherwise $\alpha_{\mathrm{LPT}}(I^c) \geqq 3$, contradicting the assumption on I^c), it must be true that $|P_i| = 1$. It follows that P_i is dominated by P_j^*, which contradicts Lemma 3.1. □

LEMMA 3.3. *No* LPT *subschedule can contain more than one task t with $l(t) \geqq \frac{3}{2}$.*

Proof. Suppose otherwise, and let P_i be the first processor to be assigned two such tasks, say t_1 and t_2. Then P_i must be the first subschedule to receive two tasks (of any length). Moreover, P_i contains exactly two tasks since $l(t_1) + l(t_2) \geqq 2(3/2) = 3$.

Let T' denote the set of the $M+1$ longest tasks of T^c. P_i must contain the two shortest tasks of T'. The optimal schedule must have assigned two elements of T' to the same processor, say P_j^*. Therefore, P_i is dominated by P_j^*, which contradicts Lemma 3.1. □

LEMMA 3.4. *In the* LPT *schedule for I^c, the length of each subschedule is at least* 8/3 *before any* LPT *subschedule length exceeds* 4.

Proof. Assume the contrary. Let y denote the first task which LPT schedules to finish after 4. Suppose P_i is the LPT subschedule containing y and let T' be the set $\{y\} \cup \{\text{all tasks scheduled before } y \text{ by LPT}\}$.

Now, the existence of some subschedule whose length is less than $\frac{8}{3}$ implies that $l(P_i - \{y\}) < \frac{8}{3}$ so that for any $t' \in T'$, $l(t') \geqq l(y) > 4 - \frac{8}{3} = \frac{4}{3}$. Therefore, $|P_i| = 2$. In fact, no LPT subschedule can contain more than two tasks of T', since y was the first task to finish after 4.

Let x represent the first task assigned to P_i. If $z \in T'$ and z is placed in a singleton subschedule in the LPT scheduling of T', then $l(z) \geqq l(x)$. Therefore the optimal subschedule that contains z cannot contain another task of T' since this subschedule would dominate P_i, contradicting Lemma 3.1. Similarly, no optimal subschedule can contain three tasks from T' since the sum of the lengths of any two tasks in T' must exceed $l(x)$.

Therefore, over T' we see that singleton subschedules for LPT and the optimal schedule are identical, and neither schedule can have subschedules with three tasks. We conclude that those tasks contained in two-task subschedules under LPT are also

in two-task subschedules of an optimal schedule. In particular, an optimal subschedule, say P_j^*, that contains x must contain another element of T'. This means that P_j^* dominates P_i, which contradicts Lemma 3.1. □

LEMMA 3.5. *If a minimal counterexample I^c exists, then there is another minimal counterexample $I^{c'}$ such that no task, t, in $T^{c'}$ satisfies $\frac{1}{3} \leqq l(t) < 1$.*

Proof. The proof is by construction. From Lemma 3.4 we see that the length of each LPT subschedule must be at least $\frac{8}{3}$ once all tasks whose lengths are longer than 1 have been assigned. Indeed, $l(P_i) \geqq 4$ for some i; this occurs as the result of the scheduling of a job with length greater than 1, at which point all the subschedules must have length at least equal to $\frac{8}{3}$. If for some t, $\frac{1}{3} \leqq l(t) < 1$, then t will be in a subschedule containing no other tasks having length less than 1. Suppose any such t is replaced by t' with $l(t') = 1$, and the LPT algorithm is rerun. Then t' will be assigned to the same subschedule which had contained t. Since t could not have been in a subschedule whose length is less than 3, the minimal subschedule length is not changed by the replacement described above. Similarly, increasing a task length in this manner cannot decrease the minimal subschedule length in the optimal schedule. Hence, the new task set T' obtained by this construction satisfies the conditions of the lemma. □

We may now assume without loss of generality that in the minimal counterexample I^c, $l(t) < \frac{1}{3}$ or $l(t) \geqq 1$ for all $t \in T^c$.

4. Proof of the main result. We are now prepared to prove that $\alpha_{\mathrm{OPT}}(I)/\alpha_{\mathrm{LPT}}(I) \leqq \frac{4}{3}$ for any problem instance. The proof will proceed by establishing that it is impossible for the counterexample I^c, described in the preceding section, to exist. The argument we use requires a weighting function $w : T^c \to \mathbb{R}^+$ to relate and compare subschedules generated by the optimal schedule and the LPT algorithm. We also need to define the set of "small" tasks contained in each subschedule. We let

$$S_i = \{t \in P_i;\ l(t) < \tfrac{1}{3}\}, \qquad 1 \leqq i \leqq M$$

and define S_j^* similarly for the subschedules of the optimal schedule.

DEFINITION 4.1 (the weighting function). Let x be any task in T^c and P_i be the LPT subschedule containing x. Then $w(x) = \frac{8}{3}$ if either $P_i - S_i = \{x\}$ or if $P_i = \{x, y\}$, $l(x) > l(y) \geqq 1$ and $l(P_i) \geqq 3$. $w(x) = \frac{4}{3}$ if $P_i = \{x, y, z\}$ and $S_i = \varnothing$, or if $P_i - S_i = \{x, y\}$ and $l(P_i - S_i) < 3$, or if $P_i = \{x, y\}$, $l(y) > l(x) \geqq 1$ and $l(P_i) \geqq 3$. $w(x) = 4l(x)/3l(S_i)$ if $x \in S_i$ and $l(P_i) \geqq 3$. Finally, $w(x) = 2l(x)/3l(S_i)$ if $x \in S_i$ and $l(P_i) < 3$.

The values of $w(\cdot)$ are summarized in Table 4.1.

TABLE 4.1
A summary of the weighting function

$w(x)$	Remarks
$\frac{8}{3}$	$P_i - S_i = \{x\}$ or $P_i = \{x, y\}$, $l(x) > l(y) \geqq 1$, and $l(P_i) \geqq 3$
$\frac{4}{3}$	$x \notin S_i$ and $w(x)$ cannot be $\frac{8}{3}$
$\dfrac{4l(x)}{3l(S_i)}$	$x \in S_i$ and $l(P_i) \geqq 3$
$\dfrac{2l(x)}{3l(S_i)}$	$x \in S_i$ and $l(P_i) < 3$

We extend w to sets of tasks T' by

$$w(T') = \sum_{x \in T'} w(x).$$

LEMMA 4.1. *Let x be a task in T^c. If $l(x) \geq 2$, then $w(x) = \frac{8}{3}$.*

PROOF. Immediate from the definition of w. □

LEMMA 4.2. *Let x be a task in T^c with $l(x) < \frac{1}{3}$. Then $w(x) > 2l(x)$.*

PROOF. Suppose to the contrary. Let P_i be the LPT subschedule containing x. If $l(P_i) \geq 3$, then $w(x) \leq 2l(x)$ implies that $2l(x) \geq 2l(x)/3l(S_i)$ or that $l(S_i) \geq \frac{2}{3}$. But this in turn implies that $l(P_i) \geq \frac{8}{3} + l(S_i) = \frac{10}{3}$ and since $l(P_i - \{x\}) < 3$, $l(x) \geq \frac{1}{3}$, which contradicts the hypothesis of the lemma.

If $l(P_i) < 3$, $w(x) \leq 2l(x)$ implies that $2l(x)/3l(S_i) \leq 2l(x)$ or that $l(S_i) \geq \frac{1}{3}$. Hence, $l(P_i) \geq \frac{8}{3} + \frac{1}{3} = 3$, another contradiction. □

LEMMA 4.3. *The weight of any* LPT *subschedule is at most* 4 *and some* LPT *subschedule has weight at most* $\frac{10}{3}$.

Proof. Let P_i be any LPT subschedule. If each of the tasks in P_i has a length at least as large as 1, then either P_i contains three tasks each having weight $\frac{4}{3}$, P_i contains a single task of weight $\frac{8}{3}$ or P_i contains two tasks, one with a weight less than or equal to $\frac{8}{3}$ and the other with a weight equal to $\frac{4}{3}$.

If P_i contains one or more tasks having length less than one, then their combined weight is at most $\frac{4}{3}$ and the combined weight of the larger tasks in P_i cannot exceed $\frac{8}{3}$.

Since there must be some subschedule whose length is less than three, its weight is either $\frac{8}{3}$ (if P_i has no task with length smaller than 1) or $\frac{10}{3}$. □

LEMMA 4.4. *The weight of any optimal subschedule is at least* 4.

Proof. Assume P_j^* is any optimal subschedule which violates the statement of the lemma. It follows from Lemma 3.2 that $|P_j^*| > 1$. Also, P_j^* must contain one or more tasks of length at least 1 or else $w(P_j^* > 2l(P_j^*) \geq 2 \cdot 4 = 8$ by Lemma 4.2. Three cases must be considered.

Case 1. Suppose P_j^* contains exactly one task x of length at least 1.

Since $l(x) < 3$, $w(S_j^*) > 2l(S_j^*) > 2 \cdot 1 = 2$ and $w(x) < 2$. But by Lemma 4.1 this means $l(x) < 2$, $l(S_j^*) > 2$ and $w(P_j^*) > 4$. Thus, the lemma holds for Case 1.

Case 2. Suppose P_j^* contains exactly two tasks x and y of length at least 1.

Each has a length smaller than 2, otherwise $w(P_j^*) \geq \frac{8}{3} + \frac{4}{3} = 4$. Since $l(P_j^*) \geq 4$, S_j^* is nonempty. It follows that $w(S_j^*) < \frac{4}{3}$, $l(S_j^*) < \frac{2}{3}$ and $l(x) + l(y) > \frac{10}{3}$. Therefore, the largest task in P_j^*, say x, has a length larger than $\frac{5}{3}$. From Lemma 3.3 we conclude that x must have been the first task assigned to some LPT subschedule, and we let z represent the second task of that subschedule. The weight of x cannot be $\frac{8}{3}$, so $l(x) + l(z) < 3$.

We choose any LPT subschedule whose length exceeds 4 (there must be at least one such subschedule) and denote its shortest task as v. Lemma 3.4 and $l(x) < 2$ indicate that z was scheduled before v and, thus, $l(z) \geq l(v)$.

Consider now any $t \in S_j^*$. Suppose $t \in P_i$, where $l(P_i) < 3$. The fact that v was not placed in P_i implies $l(v) > 4 - l(P_i - S_i) > 1 + l(S_i)$. Thus, $1 + l(S_i) + l(x) < l(v) + l(x) \leq l(z) + l(x) < 3$ and $l(S_i) < 2 - l(x)$. This means that $2l(S_i) < 4 - 2l(x) \leq 4 - l(x) - l(y) \leq l(S_j^*)$. Hence $w(t) = 2l(t)/3l(S_i) > 4l(t)/3l(S_j^*)$.

Suppose $t \in P_i$, where $l(P_i) \geq 3$. $|S_i| > 1$, otherwise $w(t) = \frac{4}{3}$, and the lemma holds. Since the length of P_i was <3 before the last task was assigned, $l(P_i) < 3 + l(S_i)/2$. Now the fact that v was not placed on P_i implies $l(v) > 4 - l(P_i - S_i) > 1 + l(S_i)/2$. Thus, using the same series of computations as performed above, we determine that $l(S_i) < l(S_j^*)$. Hence, $w(t) = 4l(t)/3l(S_i) > 4l(t)/3l(S_j^*)$.

Therefore, regardless of where the LPT algorithm placed t, $w(t) > 4l(t)/3l(S_j^*)$. But

$$w(S_j^*) = \sum_{t \in S_j^*} w(t) > \sum_{t \in S_j^*} \frac{4l(t)}{3l(S_j^*)} = \frac{4}{3l(S_j^*)} \sum_{t \in S_j^*} l(t) = \frac{4}{3}$$

and $w(P_j^*) > 4$. Thus the lemma holds for Case 2.

Case 3. Suppose P_j^* contains three or more tasks of length $\geqq 1$. Then $w(P_j^*) \geqq 3 \cdot \frac{4}{3} = 4$ and the lemma holds in any case. □

THEOREM 4.1. *For any set of M identical processors,* $Q_M(\text{LPT}) \leqq \frac{4}{3}$.

Proof. The proof follows directly from Lemmas 4.3 and 4.4. Indeed, we obtain $w(T^c) \leqq 4M - \frac{2}{3}$ from Lemma 4.3 and $w(T^c) \geqq 4M$ from Lemma 4.4. Therefore,

$$w(T^c) \leqq 4M - \tfrac{2}{3} < 4M \leqq w(T^c),$$

which is impossible. Thus, the counterexample I^c cannot exist, and the theorem is proved. □

REFERENCES

[1] E. G. COFFMAN, M. R. GAREY AND D. S. JOHNSON, *An application of bin-packing to multiprocessor scheduling*, SIAM J. Comput. 7 (1978), pp. 1–17.

[2] R. W. CONWAY, W. L. MAXWELL AND L. W. MILLER, *Theory of Scheduling*, Addison-Wesley, Reading, MA, 1967.

[3] D. K. FRIESEN AND B. L. DEUERMEYER, *Analysis of greedy solutions for a replacement part sequencing problem*, Math. Oper. Res., 6 (1981), pp. 74–87.

[4] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability*, Freeman, San Francisco, 1979.

[5] R. L. GRAHAM, *Bounds on multiprocessor timing anomalies*, SIAM J. Appl. Math, 17 (1969), pp. 416–429.

[6] S. C. PARIKH, *On a fleet sizing and allocation problem*, Management Sci., 23 (1977), pp. 972–977.

[7] A. H. G. RINNOOY-KAN, *Machine Scheduling Problems*, Nijhoff, The Hague, 1976.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 2, June 1982

0196-5212/82/0302-0007 $01.00/0

NONNEGATIVE MATRICES A SUCH THAT $Ax = b$ HAS NONNEGATIVE BEST APPROXIMATE SOLUTION*

YOSHIMI EGAWA† AND S. K. JAIN‡

Abstract. In this paper the question of existence of nonnegative best approximate solutions (b.a.s.) of the linear system $Ax = b$ is investigated. Firstly, a necessary condition that $Ax = b$ have a nonnegative b.a.s. for all $b \geqq 0$ with respect to a positive definite symmetric bilinear form S whose associated matrix is nonnegative is obtained. It follows as a consequence that $Ax = b$ also has a nonnegative least squares solution (l.s.s.). Among other results it is proved that if B is a nonnegative idempotent matrix such that $AB = BA$, rank (AB) = rank A, then $Ax = b$ has a nonnegative l.s.s. for all $b \in R(B)$, $b \geqq 0$, if and only if for certain well-defined matrix A_0 (called the coefficient matrix of A with respect to B) and certain symmetric bilinear form S, $A_0 x = b$ has a nonnegative b.a.s. with respect to S. These results generalize the well-known results concerning the question of the existence of a nonnegative l.s.s. for the system $Ax = b$. Indeed, these investigations initiate a new approach to the question beyond the technique of inverse-positivity. The importance of this question lies in its varied applications to problems in mathematical economics, in probability theory, in operations research and in numerical algebra.

1. Notation and definitions.

$\mathbb{R}^m$: the vector space of $m \times 1$ matrices over the reals $\mathbb{R}$.
$\|x\|_2$: the usual Euclidean norm of a vector x.
$(X)^i$: the ith column of a matrix X.
$(X)_i$: the ith row of a matrix X.
$(x)_i$: the ith entry of a vector x.
$X_{i,j}$: the (i, j)th entry of a matrix X; thus, $((X)^j)_i = X_{i,j}$.
X^t: the transpose of a matrix X.
$R(B)$: the range of an $m \times n$ matrix B, i.e., $\{y \in \mathbb{R}^m | y = Bx$, for some $x \in \mathbb{R}^n\}$.
$R(B)^{\perp s}$: the subspace of $\mathbb{R}^m$ consisting of vectors x such that $S(x, b) = 0$ for all $b \in R(B)$, where S is a positive definite symmetric bilinear form.
$R(B)^{\perp}$: the subspace $R(B)^{\perp s}$ when S is the usual inner product on $\mathbb{R}^m$.
$\langle Y \rangle$: the subspace spanned by the subset Y of $\mathbb{R}^m$.
e_i: the vector in $\mathbb{R}^m$ having all entries zero except the ith entry, which is 1.
0_m: the zero vector in $\mathbb{R}^m$.

Let $S(\cdot, \cdot)$ be a positive definite symmetric bilinear form over $\mathbb{R}^m$. Then the associated symmetric matrix with respect to the standard basis $(e_1, e_2, \ldots, e_m)$, shall also be denoted by S, i.e., $S(x, y) = x^t S y$, $x, y \in \mathbb{R}^m$.

Let A be an $m \times n$ matrix and let $b \in \mathbb{R}^m$. Then $x_0 \in \mathbb{R}^n$ is called a best approximate solution of the system $Ax = b$ with respect to (w.r.t.) S if $S(Ax_0 - b, Ax_0 - b)$ is minimum. If S is the usual Euclidean norm, then the best approximate solution with respect to S is commonly known as the least squares solution.

If A, X are respectively $m \times n$, $n \times m$ matrices such that $AXA = A$ and $(AX)^t = AX$, then X is called a $\{1, 3\}$-inverse of A and is denoted by $A^{(1,3)}$.

For simplicity, we shall not indicate the order of matrices if it is clear from the context. Further, all matrices are real.

* Received by the editors March 1, 1981, and in final revised form September 2, 1981.
† Department of Mathematics, Ohio State University, Columbus, Ohio 43210.
‡ Department of Mathematics, Ohio University, Athens, Ohio, 45701. The work of this author was done while visiting Ohio State University during the Fall of 1980.

2. Introduction. This paper addresses the question of characterizing nonnegative matrices A such that the linear system $Ax = b$, for certain nonnegative vectors b, has a nonnegative best approximate solution. The importance of this question can hardly be overemphasized in view of the fact that in many of the applications of nonnegative matrices one is involved in finding nonnegative solutions or least squares solutions of the system $Ax = b$, where $A \geqq 0$, $b \geqq 0$. For example, one finds numerous applications in areas such as mathematical economics, probability theory, numerical algebra and linear programming.

Since the nonnegativity of certain generalized inverses of A is related to the existence of nonnegative least squares solution of the system $Ax = b$, many authors have previously considered this question from this viewpoint. For example, the existence of a nonnegative $\{1, 3\}$-inverse of A is equivalent to the existence of a nonnegative least squares solution of $Ax = b$ for all nonnegative vectors b. The characterization of nonnegative matrices A having a certain nonnegative generalized inverse has been extensively studied in the literature (see [1]–[3], [6]–[11]).

We begin by considering the nonnegative best approximate solutions of $Ax = b$ with respect to an arbitrary positive definite symmetric bilinear form S whose associated matrix is nonnegative, and show in Theorem 3.7 that $A^{(1,3)} \geqq 0$—a well-known result for the Euclidean norm. We then proceed to the main question addressed in this paper, that of characterizing nonnegative matrices A such that $Ax = b$ has a nonnegative least squares solution for all nonnegative vectors b in a given set. We study this question in the case when $b \in R(B)$, where B is a nonnegative idempotent matrix such that $AB = BA$, $\operatorname{rank}(AB) = \operatorname{rank} A$ (Theorem 4.4). This is done by first obtaining the characterization of nonnegative matrices A which commute with a given nonnegative idempotent matrix B such that $\operatorname{rank}(AB) = \operatorname{rank} A$ (Lemma 4.2). We then introduce an intrinsic matrix A_0 (coefficient matrix) of A. The problem of the nonnegative least squares solution of $Ax = b$, $b \in R(B)$, $b \geqq 0$, is then reduced to the problem of obtaining a nonnegative best approximate solution of $A_0 x = b$, for all nonnegative vectors b, with respect to some suitably defined norm S (Theorem 4.3). The proof of Theorem 4.4 is then completed by applying Theorem 3.7 and Theorem 4.3. An example is given to show that the converse of Theorem 4.4 does not, in general, hold.

We emphasize that Theorem 4.4 is an initial attempt to study the question stated in the beginning of the introduction. That this theorem is also true under a certain weaker hypothesis is explained in Remark 3 at the end of the paper. However, it is desirable that the hypothesis in Theorem 4.4 be further weakened. This remains open.

We remark that Lemmas 3.1, 3.6 and 4.2 are also of independent interest.

3. Nonnegative best approximate solutions.

LEMMA 3.1. *Let A be a nonnegative $m \times n$ matrix of rank r. Suppose $Ax = b$ has a nonnegative solution for every $b \geqq 0$, which makes this equation consistent. Then there exist permutation matrices P, Q such that*

$$(PAQ)_{i,i} \neq 0, \qquad 1 \leqq i \leqq r,$$

$$(PAQ)_{i,j} = 0, \qquad 1 \leqq i < j \leqq r.$$

Proof. We proceed by induction on r. Let $\mathcal{Q}$ denote the set of ordered pairs (P, Q) of permutation matrices such that $(PAQ)^j$, $1 \leqq j \leqq r$, are linearly independent, and $(PAQ)_{1,1} \neq 0$. For each $(P, Q) \in \mathcal{Q}$, we define $q(P, Q)$ as follows:

$$q(P, Q) = \operatorname{card}\{j \mid 1 \leqq j \leqq r, (PAQ)_{1,j} \neq 0\}.$$

Let

$$p = \min \{q(P, Q) | (P, Q) \in \mathcal{Q}\}.$$

We want to prove $p = 1$. Suppose $p \geqq 2$. Let

$$\mathcal{F} = \{(P, Q) \in \mathcal{Q} | q(P, Q) = p, (PAQ)_{1,2} \neq 0\}.$$

Let

$$\alpha = \min \{(PAQ)_{1,1}/(PAQ)_{1,2} | (P, Q) \in \mathcal{F}\}.$$

Suppose (P, Q) is an element of $\mathcal{F}$ which gives this minimum value α. Set $F = PAQ$. Let

$$L = \{j | 1 \leqq j \leqq r, F_{1,j} = 0\},$$

$$K = \{i | 1 \leqq i \leqq m, \alpha F_{i,2} > F_{i,1}, F_{i,j} = 0 \ \forall j \in L\}.$$

Suppose $K \neq \varnothing$ and $i \in K$. Then, if we replace the first row by the ith row, this contradicts the minimality of α unless $F_{i,j} = 0$ for some $j \notin L$, $1 \leqq j \leqq r$. In the latter case, we get a contradiction to the minimality of p. Therefore $K = \varnothing$. This implies there exist nonnegative numbers β_j's, $j \in L$ such that

$$b = (F)^1 - \alpha(F)^2 + \sum_{j \in L} \beta_j (F)^j \geqq 0.$$

By our assumption the system $Fx = b$ has a nonnegative solution. Thus there exist $\gamma_j \geqq 0$, $1 \leqq j \leqq n$, such that $b = \sum_{j=1}^{n} \gamma_j (F)^j$. By the replacement theorem, we can choose k with $\gamma_k \neq 0$ such that

$$\langle (F)^j | 1 \leqq j \leqq r \rangle = \langle (F)^k, (F)^j | 1 \leqq j \leqq r, j \neq 2 \rangle.$$

Since $(b)_1 = 0$ and $\gamma_k \neq 0$, we have $F_{1,k} = 0$. Hence, if we replace the second column by the kth column, we get a contradiction to the minimality of p. Thus $p = 1$. By interchanging rows and columns suitably, we may assume that there exists $l \geqq r$ such that

$$A_{1,j} = 0, \quad 2 \leqq j \leqq l, \qquad A_{1,j} \neq 0, \quad j \geqq l+1 \text{ and } j = 1,$$

and the submatrix A' of A which consists of the columns 2 through l of A is of rank $r-1$. One can check that A' satisfies the hypothesis of the lemma. Consider the submatrix A'_0 consisting of all but the first row of A'. Since the first row of A' is a zero vector, we may assume by applying induction to A'_0 that

$$A_{i,j} = 0, \quad 2 \leqq i < j \leqq r, \qquad A_{i,i} \neq 0, \quad 2 \leqq i \leqq r.$$

Since $A_{1,j} = 0$, $2 \leqq j \leqq r$, and $A_{1,1} \neq 0$, the proof is complete.

In our next lemma, we shall need the following notation. Let $u \in \mathbb{R}^m$. We define

$$Z(u) = \{i | 1 \leqq i \leqq m, (u)_i \neq 0\},$$

$$Z_1(u) = \{i \in Z(u) | 1 \leqq i \leqq r\}.$$

LEMMA 3.2. *Let A be an $m \times n$ nonnegative matrix of rank r. Suppose that for every integer l_1, $1 \leqq l_1 \leqq n$, and for every subset L of $\{1, 2, \cdots, n\}$ with $Z_1((A)^{l_1}) \subseteq Z_1(\sum_{l \in L} (A)^l)$ we have the inclusion $Z((A)^{l_1}) \subseteq Z(\sum_{l \in L} (A)^l)$. (If L is empty, then by $\sum_{l \in L} (A)^l$ we understand the zero vector.) Also suppose that $Ax = b$ has a nonnegative solution for every $b \geqq 0$ which makes this equation consistent. Then there exist permutation*

matrices P, Q such that

$$P = \begin{bmatrix} P_1 & 0 \\ 0 & I \end{bmatrix},$$

where P_1 is a permutation matrix of order r and

$$(PAQ)_{i,i} \neq 0, \qquad 1 \leqq i \leqq r,$$

$$(PAQ)_{i,j} = 0, \qquad 1 \leqq i \leqq r, \quad 1 \leqq j \leqq r, \quad i \neq j.$$

Proof. Let P', Q' be permutation matrices which satisfy the conclusion of Lemma 3.1. Now, since $(P'AQ')_{j,j} \neq 0$ and $(P'AQ')_{j,l} = 0$ for $j < l \leqq r$, we have

$$Z\Big(\sum_{l=j+1}^{r} (P'AQ')^l\Big) \not\supseteq Z((P'AQ')^j), \qquad 1 \leqq j \leqq r.$$

Therefore,

$$Z\Big(\sum_{l=j+1}^{r} (AQ')^l\Big) \not\supseteq Z((AQ')^j), \qquad 1 \leqq j \leqq r.$$

But then by assumption

$$Z_1\Big(\sum_{l=j+1}^{r} (AQ')^l\Big) \not\supseteq Z_1((AQ')^j),$$

and so by choosing $j = r, r-1, \ldots, 1$, we obtain

$$1 \leqq \operatorname{card}(Z_1((AQ')^r)) \neq \operatorname{card}\left(Z_1\Big(\sum_{l=r-1}^{r} (AQ')^l\Big)\right)$$

$$\neq \cdots \neq \operatorname{card}\left(Z_1\Big(\sum_{l=1}^{r} (AQ')^l\Big)\right) \leqq r.$$

Hence there exists a permutation matrix P satisfying the following property:

(*) $$(PAQ')_{i,i} \neq 0, \quad 1 \leqq i \leqq r \qquad (PAQ')_{i,j} = 0, \quad 1 \leqq i < j \leqq r,$$

where

$$P = \begin{bmatrix} P_1 & 0 \\ 0 & I \end{bmatrix};$$

P_1 is a permutation matrix of order r. With the matrix P as obtained above, let

$$\mathscr{B} = \{Q, \text{ a permutation matrix} \mid PAQ \text{ has the property } (*)\}.$$

By way of contradiction, suppose Lemma 3.2 is false. For each $Q \in \mathscr{B}$, let $\lambda(Q)$ be the positive integer defined as follows:

$$\lambda(Q) = \max\{i \mid 2 \leqq i \leqq r \text{ such that } \exists j,\ 1 \leqq j \leqq i-1, \text{ with } (PAQ)_{i,j} \neq 0\}.$$

Let

$$q = \min\{\lambda(Q) \mid Q \in \mathscr{B}\}, \qquad \mathscr{C} = \{Q \in \mathscr{B} \mid \lambda(Q) = q\}.$$

Let

$$\alpha = \min\left\{\sum_{1 \leqq j \leqq q-1} \frac{(PAQ)_{q,j}}{(PAQ)_{j,j}} \,\middle|\, Q \in \mathscr{C}\right\}.$$

Now let Q be an element of $\mathscr{C}$ which gives the minimum value α. Set $F = PAQ$. We know there exists j_0, $1 \leqq j_0 \leqq q-1$, such that $F_{q,j_0} \neq 0$. By the maximality of $\lambda(Q)$,

$$F_{i,q} = 0, \qquad i \neq q, \quad 1 \leqq i \leqq r.$$

Therefore $Z_1((F)^q) \subseteq Z_1((F)^{j_0})$. Since F satisfies the hypothesis of Lemma 3.2, $Z((F)^q) \subseteq Z((F)^{j_0})$. Therefore, there exists $\beta > 0$ such that $(F)^{j_0} - \beta(F)^q \geqq 0$. Let us set

$$f = (F)^{j_0} - \beta(F)^q. \tag{1}$$

Then $Fx = f$ must have a nonnegative solution. So we may write

$$f = \sum_{j=1}^{n} \gamma_j (F)^j, \qquad \gamma_j \geqq 0. \tag{2}$$

It follows from (2) that there exists j_1, $1 \leqq j_1 \leqq n$, such that

$$\gamma_{j_1} \neq 0, \quad ((F)^{j_1})_{j_0} \neq 0, \quad \frac{((F)^{j_1})_q}{((F)^{j_1})_{j_0}} \leqq \frac{(f)_q}{(f)_{j_0}}. \tag{3}$$

On the other hand, from (1) we have $(f)_q < ((F)^{j_0})_q$ and $(f)_{j_0} = ((F)^{j_0})_{j_0}$, and so

$$\frac{(f)_q}{(f)_{j_0}} < \frac{((F)^{j_0})_q}{((F)^{j_0})_{j_0}}. \tag{4}$$

Again, by (1),

$$(f)_i = 0, \qquad 1 \leqq i \leqq j_0 - 1, \quad q+1 \leqq i \leqq r. \tag{5}$$

Thus by (2) and (5), and the fact that $\gamma_{j_1} \neq 0$, we obtain

$$((F)^{j_1})_i = 0, \qquad 1 \leqq i \leqq j_0 - 1, \quad q+1 \leqq i \leqq r. \tag{6}$$

Therefore, if we replace $(F)^{j_0}$ by $(F)^{j_1}$, we shall get a smaller value of α by (3), (4) and (6) except when $((F)^{j_1})_q = 0$ and $F_{q,j} = 0$, $1 \leqq j \leqq q-1$, $j \neq j_0$. In the latter case we get a smaller value of q. Thus in each case we arrive at a contradiction. This completes the proof.

SUBLEMMA 3.3. *Let A be an $m \times n$ matrix of rank r. Let S be a positive definite symmetric bilinear form over $\mathbb{R}^m$. Then there exists a subset Λ of cardinality r of $\{1, 2, \cdots, m\}$ such that*

$$\langle e_i | i \in \Lambda \rangle \cap R(A)^{\perp s} = 0$$

and

$$\forall i \in \Lambda, \quad \exists j, \quad 1 \leqq j \leqq n, \quad A_{i,j} \neq 0.$$

Proof. Let

$$\Delta = \{i | 1 \leqq i \leqq m, \exists j, 1 \leqq j \leqq n, \text{ such that } A_{i,j} \neq 0\}.$$

Then

$$R(A) \subseteq \langle e_i | i \in \Delta \rangle$$

and so

$$\dim (R(A)^{\perp s} \cap \langle e_i | i \in \Delta \rangle) = (\text{card } \Delta) - r.$$

By choosing Λ to be a maximal subset of Δ such that $\langle e_i | i \in \Lambda \rangle \cap R(A)^{\perp s} = 0$, we get the desired conclusion.

We now state without proof some basic facts contained in the following two sublemmas.

SUBLEMMA 3.4. *Let A be an $m \times n$ matrix, and let b be a vector of size m. Let S be a positive definite symmetric bilinear form over $\mathbb{R}^m$. Let $b = b_1 + b_2$, $b_1 \in R(A)$, $b_2 \in R(A)^{\perp_S}$. Then x_0 is a best approximate solution of the system $Ax = b$ with respect to S if and only if $Ax_0 = b_1$.*

SUBLEMMA 3.5. *Let A be an $m \times n$ matrix, and b be a vector of size m. Let S be a positive definite symmetric bilinear form over $\mathbb{R}^m$. Let P, Q be permutation matrices of orders m, n respectively. Then x_0 is a best approximate solution of the system $Ax = b$ with respect to S if and only if $Q^{-1}x_0$ is a best approximate solution of the system $(PAQ)x = Pb$ with respect to PSP^{-1}.*

LEMMA 3.6. *Let A be an $m \times n$ nonnegative matrix of rank r. Let S be a positive definite symmetric bilinear form over $\mathbb{R}^m$. Suppose that $Ax = b$ has a nonnegative best approximate solution with respect to S for every $b \geqq 0$.*

Then there exist permutation matrices P, Q such that

$$(PAQ)_{i,i} \neq 0, \qquad 1 \leqq i \leqq r,$$

$$(PAQ)_{i,j} = 0, \qquad 1 \leqq i \leqq r, \quad 1 \leqq j \leqq r, \quad i \neq j,$$

$$\langle e_i | 1 \leqq i \leqq r \rangle \cap R(PAQ)^{\perp_{PSP^{-1}}} = 0.$$

Proof. Let Λ be as in the conclusion of Sublemma 3.3. Without any loss of generality, we may assume $\Lambda = \{1, 2, \cdots, r\}$. Clearly, for each $k \notin \Lambda$, $1 \leqq k \leqq m$, there exists a unique vector q_k of $R(A)^{\perp_S}$ such that

$$(q_k)_i = \begin{cases} 1, & i = k, \\ 0, & i \neq k, \end{cases} \qquad r+1 \leqq i \leqq m.$$

In order to prove our lemma, it suffices to prove that A satisfies the hypothesis of Lemma 3.2. Let $Z_1(u)$, $Z(u)$ be as in Lemma 3.2. By way of contradiction, let $(A)^{l_1}$ and $a = \sum_{l \in L} (A)^l$ be such that

$$Z_1((A)^{l_1}) \subseteq Z_1(a) \quad \text{but} \quad Z((A)^{l_1}) \not\subseteq Z(a).$$

We choose $z \in R(A)$ such that $q_k + z \geqq 0$ for all $k \in \{r+1, \cdots, m\}$. Further, for each vector $u \in \mathbb{R}^m$, let

$$T(u) = \{k | r+1 \leqq k \leqq m, k \notin Z(a), (u)_k < 0\}.$$

Assume $T(u) \neq \varnothing$. Let us set

$$p(u) = \min T(u).$$

Next we choose positive number $\alpha(u)$ such that

$$(u + \alpha(u)(q_{p(u)} + z))_{p(u)} = 0.$$

Now let $u_0 = -(A)^{l_1}$, and define v_i, w_i, u_i inductively by

$$v_i = \alpha(u_{i-1}) q_{p(u_{i-1})}, \quad w_i = \alpha(u_{i-1}) z, \quad u_i = u_{i-1} + v_i + w_i.$$

We continue until $T(u_t) = \varnothing$ for some positive integer t. For each $i = 1, 2, \cdots, t$, we have

$$T(u_i) \subset T(u_{i-1}),$$
$$(u_i)_{p(u_{i-1})} = 0,$$
$$(u_{i-1} + w_i)_{p(u_{i-1})} < 0,$$
$$(u_i)_k \geqq 0, \qquad 1 \leqq k \leqq r, \quad k \notin Z_1(a).$$

Let $v = \sum_{i=1}^{t} v_i$, $w = u_0 + \sum_{i=1}^{t} w_i$ and $u = u_t$. Let us also write $p = p(u_{t-1})$ for convenience. Then

$$u = v + w, \qquad v \in R(A)^{\perp_S}, \quad w \in R(A),$$
$$(u)_k \geqq 0 \quad \text{for all } k \notin Z(a), \quad 1 \leqq k \leqq m,$$
$$(w)_p < 0,$$

By the definition of $Z(a)$, there exists $\beta > 0$ such that $\beta a + u \geqq 0$. Further, since $p \notin Z(a)$, $(\beta a + w)_p = (w)_p < 0$. This implies that $Ax = \beta a + u$ does not have any nonnegative best approximate solution with respect to the norm S, a contradiction. Hence A satisfies the hypothesis of Lemma 3.2, completing the proof.

THEOREM 3.7. *Let A be an $m \times n$ matrix of rank r. Let S be a positive definite symmetric bilinear form over $\mathbb{R}^m$ satisfying*

$$(**) \qquad S(e_i, e_k) \geqq 0, \qquad 1 \leqq i \leqq m, \quad 1 \leqq k \leqq m.$$

Suppose $Ax = b$ has a nonnegative best approximate solution with respect to S for every $b \geqq 0$. Then there exist permutation matrices P, Q such that

$$PAQ = \begin{bmatrix} J & JD \\ 0 & 0 \end{bmatrix},$$

where

$$J = \begin{bmatrix} z_1 & & & \\ & z_2 & & \\ & & \ddots & \\ & & & z_r \end{bmatrix},$$

z_i is a positive vector of size λ_i and D is some nonnegative matrix; or equivalently A has a nonnegative $\{1, 3\}$-inverse. (The zero block row in the description of PAQ may be absent.)

Proof. By Lemma 3.6 there exist permutation matrices P, Q such that

$$(PAQ)_{i,i} \neq 0, \qquad 1 \leqq i \leqq r,$$
$$(PAQ)_{i,j} = 0, \qquad 1 \leqq i, j \leqq r, \quad i \neq j$$

and that

$$\langle e_i | 1 \leqq i \leqq r \rangle \cap R(PAQ)^{\perp_{PSP^{-1}}} = 0.$$

For each $r + 1 \leqq k \leqq m$, we define q_k to be the unique vector in $R(PAQ)^{\perp_{PSP^{-1}}}$ such that

$$(q_k)_k = 1, \qquad (q_k)_i = 0, \quad i \neq k, \quad r + 1 \leqq i \leqq m.$$

For each $k \in \{r + 1, \cdots, m\}$, let

$$p_k = \text{card}\,\{j | 1 \leqq j \leqq r, (PAQ)_{k,j} \neq 0\}.$$

We have only to show that $p_k \leqq 1$, $r+1 \leqq k \leqq m$. By way of contradiction, suppose $p_{k_0} \geqq 2$ for some k_0. By (**), there exists i_0, $1 \leqq i_0 \leqq r$, such that

$$(q_{k_0})_{i_0} < 0.$$

Clearly, there exist nonnegative numbers α_j's, $r+1 \leqq j \leqq m$, $j \neq k_0$, such that

$$\left(-(PAQ)^{i_0} + \sum_{\substack{r+1 \leqq j \leqq m \\ j \neq k_0}} \alpha_j q_j\right)_k \geqq 0$$

for all $k \in \{r+1, \cdots, m\}$, $k \neq k_0$. Since $(q_{k_0})_{i_0} < 0$, there exists $\alpha_{k_0} < 0$ such that

$$\left(-(PAQ)^{i_0} + \sum_{r+1 \leqq j \leqq m} \alpha_j q_j\right)_k \geqq 0$$

for all $k \in \{r+1, r+2, \cdots, m\}$ with $k \neq k_0$ and for $k = i_0$. Since $p_{k_0} \geqq 2$, there exists j_0, $1 \leqq j_0 \leqq r$, $j_0 \neq i_0$ such that $(PAQ)_{k_0, j_0} \neq 0$. Then there exists $\beta > 0$ such that

$$\left(-(PAQ)^{i_0} + \beta(PAQ)^{j_0} + \sum_{r+1 \leqq j \leqq m} \alpha_j q_j\right)_k \geqq 0$$

for all $k \in \{r+1, r+2, \cdots, m\}$ and for $k = i_0$. Finally, there exist nonnegative numbers γ_j's, $1 \leqq j \leqq r$, $j \neq i_0$, such that

$$b = -(PAQ)^{i_0} + \beta(PAQ)^{j_0} + \sum_{\substack{1 \leqq j \leqq r \\ j \neq i_0}} \gamma_j (PAQ)^j + \sum_{r+1 \leqq j \leqq m} \alpha_j q_j \geqq 0.$$

Since

$$\left(b - \sum_{r+1 \leqq j \leqq m} \alpha_j q_j\right)_{i_0} = -(PAQ)_{i_0, i_0} < 0,$$

the equation $PAQx = b$ does not have any nonnegative best approximate solution with respect to the bilinear form PSP^{-1}. Hence, by Sublemma 3.5 the system $Ax = P^{-1}b$ does not have any nonnegative best approximate solution with respect to S, a contradiction. This gives us the desired structure of A. The last statement follows from the theorem of Berman–Plemmons [3, Thm. 5].

We now proceed to give certain remarks about sufficiency conditions in order that for all b, $Ax = b$ have a nonnegative best approximate solution.

Remark 3.8. Let A be a nonnegative $m \times n$ matrix of the form

$$A = \begin{bmatrix} J & JD \\ 0 & 0 \end{bmatrix},$$

where J and D are as in the statement of Theorem 3.7. Let S be a positive definite symmetric bilinear form satisfying $S(e_i, e_k) \geqq 0$, $1 \leqq i, k \leqq m$. Then the following two statements are equivalent:

(i) $Ax = b$ has a nonnegative best approximate solution w.r.t. S for all nonnegative vectors $b \in \mathbb{R}^m$.

(ii) For each $v \in R(A)^{\perp s}$, either there exists $k \geqq (\sum_{i=1}^{r} \lambda_i) + 1$ such that $(v)_k < 0$ or

$$\forall j, \quad 1 \leqq j \leqq r, \quad \exists k_j \quad \text{with} \left(\sum_{i=1}^{j-1} \lambda_i\right) + 1 \leqq k_j \leqq \sum_{i=1}^{j} \lambda_i$$

such that $(v)_{k_j} \leqq 0$.

Proof. Straightforward.

Remark 3.9. Condition (ii) in the above remark is automatically satisfied if S is diagonal. Thus, for such an S, the converse of Theorem 3.7 also holds.

4. Nonnegative least squares solution. The characterization of nonnegative idempotent matrices plays an important role in this section. We state this in the following lemma due to Flor [4].

LEMMA 4.1. [4, Thm. 2]. *Let B be a nonnegative idempotent matrix of rank s. Then there exists a permutation matrix P such that*

$$PBP^t = \begin{bmatrix} J & JD & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CJ & CJD & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where J is a direct sum of matrices $x_i y_i^t$, where x_i, y_i are positive vectors such that $y_i^t x_i = 1$, $1 \leqq i \leqq s$ and C, D are nonnegative matrices of suitable sizes.

The lemma that follows characterizes all real matrices A which commute with a nonnegative idempotent matrix B such that rank AB = rank A.

LEMMA 4.2. *Let B be an idempotent matrix of rank s of the form*

$$\begin{bmatrix} J & JD & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CJ & CJD & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where diagonal blocks are square matrices of orders a_1, a_2, a_3, a_4, J is a direct sum of $m_i \times m_i$ matrices $x_i y_i^t$ with $y_i^t x_i = 1$ and x_i, y_i having no zero entry, $1 \leqq i \leqq s$. Let A be a square matrix such that $AB = BA$ and rank AB = rank A. *Then*

$$A = \begin{bmatrix} K & KD & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CK & CKD & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where the diagonal blocks are square matrices of orders a_1, a_2, a_3, a_4 and $K = (K_{ij})$, $1 \leqq i, j \leqq s$ where $K_{ij} = \beta_{ij} x_i y_j^t$ is an $m_i \times m_j$ block matrix.

Furthermore, $AB = A = BA$.

Proof. Let $x_{i,j}$ and $y_{i,j}$ denote the ith entry of x_j and y_j respectively. Set

$$n_i = \sum_{j=1}^{i-1} m_j, \quad 2 \leqq i \leqq s, \qquad n_1 = 0.$$

Then

$$(B)^{n_i+l_1} = \frac{y_{l_1,i}}{y_{l_2,i}} (B)^{n_i+l_2}, \qquad 1 \leqq i \leqq s, \quad 1 \leqq l_1, l_2 \leqq m_i.$$

Therefore,

$$(AB)^{n_i+l_1} = \frac{y_{l_1,i}}{y_{l_2,i}} (AB)^{n_i+l_2}, \qquad 1 \leqq i \leqq s, \quad 1 \leqq l_1, l_2 \leqq m_i. \tag{7}$$

Now, we have

$$(B)^l = \sum_{k=1}^{a_1} d_{kl}(B)^k, \tag{8}$$

where

$$d_{kl} = \begin{cases} \delta_{kl}, & l \leqq a_1, \\ (k, l-a_1)\text{-entry of } D, & a_1+1 \leqq l \leqq a_1+a_2, \\ 0, & l \geqq a_1+a_2+1. \end{cases}$$

It follows from (8) that

$$(AB)^l = \sum_{k=1}^{a_1} d_{kl}(AB)^k. \tag{9}$$

By (7) and (9), we obtain

$$R(AB) = \langle (AB)^{n_i+1} | 1 \leqq i \leqq s \rangle. \tag{10}$$

Let $r = \text{rank } (AB)$. Since $AB = BA$, (10) implies that we can choose r linearly independent vectors among $(BA)^{n_i+1}$, $1 \leqq i \leqq s$. By simultaneous rearrangement of rows and columns, we may assume that $(BA)^{n_i+1}$, $1 \leqq i \leqq r$, are linearly independent. Then since $(BA)^l = B(A)^l$, we get that $(A)^{n_i+1}$, $1 \leqq i \leqq r$, are linearly independent and, hence, form a basis of $R(A)$. Therefore an arbitrary column $(A)^l$ of A can be expressed as

$$(A)^l = \sum_{i=1}^{r} \alpha_{li}(A)^{n_i+1}. \tag{11}$$

Then

$$(BA)^l = \sum_{i=1}^{r} \alpha_{li}(BA)^{n_i+1}. \tag{12}$$

From (7) and (12),

$$\alpha_{n_j+l_1,i} = \frac{y_{l_1,j}}{y_{l_2,j}} \alpha_{n_j+l_2,i}.$$

The above together with (11) yields

$$(A)^{n_j+l_1} = \frac{y_{l_1,j}}{y_{l_2,j}} (A)^{n_j+l_2}. \tag{13}$$

Similarly,

$$(A)_{n_j+l_1} = \frac{x_{l_1,j}}{x_{l_2,j}} (A)_{n_j+l_2}. \tag{14}$$

Set

$$\beta_{ij} = \frac{(n_i+1, n_j+1)\text{-entry of } A}{x_{1,i}y_{1,j}}.$$

This gives the desired structure for K. Now let $l \geqq a_1 + 1$. Then

$$(BA)^l = \sum_{k=1}^{a_1} d_{kl}(BA)^k \qquad \text{(from (9))}$$

$$= \sum_{k=1}^{a_1} d_{kl} \sum_{i=1}^{r} \alpha_{ki}(BA)^{n_i+1} \qquad \text{(from (12))}$$

$$= \sum_{i=1}^{r} \left(\sum_{k=1}^{a_1} d_{kl}\alpha_{ki} \right)(BA)^{n_i+1}.$$

Thus from (12)

$$\alpha_{li} = \sum_{k=1}^{a_1} d_{kl}\alpha_{ki}.$$

Hence by (11)

$$(A)^l = \sum_{i=1}^{r} \left(\sum_{k=1}^{a_1} d_{kl}\alpha_{ki} \right)(A)^{n_i+1} = \sum_{k=1}^{a_1} d_{kl} \left(\sum_{i=1}^{r} \alpha_{ki}(A)^{n_i+1} \right)$$

$$= \sum_{k=1}^{a_1} d_{kl}(A)^k \quad \text{(from (11))}.$$

Let X be the submatrix of A consisting of its first a_1 columns. Then by the above equation, and by the definition of d_{kl}, we obtain

$$A = [X \quad XD \quad 0 \quad 0].$$

A similar argument for rows yields

$$X = \begin{bmatrix} K \\ 0 \\ CK \\ 0 \end{bmatrix}.$$

Hence A is of the desired form. The last statement is obvious. This completes the proof.

The $s \times s$ matrix (β_{ij}) in the above lemma will be referred to as a coefficient matrix of A with respect to B. More generally, let B be an arbitrary nonnegative idempotent matrix, and let P be a permutation matrix such that PBP^t is as in Lemma 4.1. Let A be a nonnegative matrix such that $AB = BA$ and rank (AB) = rank A. Then we can define a coefficient matrix (β_{ij}) of PAP^t with respect to PBP^t. We refer to this matrix (β_{ij}) also as a coefficient matrix of A with respect to B. We remark that this definition of coefficient matrix of A is unique up to similarity by a monomial matrix. For, if A, B and P are as above and if we write

$$PBP^t = U'_P V'^t_P,$$

where

$$U'_P = \begin{bmatrix} U \\ 0 \\ CU \\ 0 \end{bmatrix}, \qquad U_P = \begin{bmatrix} x_1 & & & \\ & x_2 & & \\ & & \ddots & \\ & & & x_s \end{bmatrix},$$

$$V'_P = \begin{bmatrix} V \\ D^tV \\ 0 \\ 0 \end{bmatrix}, \qquad V_P = \begin{bmatrix} y_1 & & & \\ & y_2 & & \\ & & \ddots & \\ & & & y_s \end{bmatrix},$$

then the coefficient matrix defined above is the $s \times s$ matrix A_P such that $PAP^t = U'_P A_P V'^t_P$. Note that U_P and V_P are not unique even if P is fixed, but that if P, U_P, V_P are all fixed, then A_p is uniquely determined. Now suppose that for some permutation matrix Q, $QBQ^t = U'_Q V'^t_Q$ is also as in Lemma 4.1. Then the matrix A_Q such that $QAQ^t = U'_Q A_Q V'^t_Q$ is obtained as follows:

$$\begin{aligned} QAQ^t &= QP^{-1}U'_P A_P V'^t_P (QP^{-1})^t \\ &= (QP^{-1}U'_P Q_0^{-1})(Q_0 A_P Q_0^{-1})(Q_0 V'^t_P (QP^{-1})^t), \end{aligned}$$

where Q_0 is the unique $s \times s$ monomial matrix such that $QP^{-1}U'_P Q_0^{-1} = U'_Q$ (or equivalently, $QP^{-1}V'_P Q_0^t = V'_Q$.) Thus,

$$A_Q = Q_0 A_P Q_0^{-1}.$$

This justifies our remark that the coefficient matrix is determined up to similarity by a monomial matrix. Before proving our next main result, we fix the following notation.

Let B be a nonnegative idempotent matrix of rank s. We shall without any loss of generality assume (by using Lemma 4.1) that B is of the form

$$\begin{bmatrix} J & JD & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CJ & CJD & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where J is a matrix as in Lemma 4.1. Let $C = (c_{ij})$, $1 \leqq i \leqq a_3$, $1 \leqq j \leqq a_1$. Let $x_{i,j}$, $y_{i,j}$, m_i, n_j be as in proof of Lemma 4.2. Set

$$g_{jk} = \sum_{i=1}^{m_k} (c_{j,n_k+i} x_{i,k}), \qquad 1 \leqq j \leqq a_3, \quad 1 \leqq k \leqq s,$$

$$h_{kl} = \sum_{j=1}^{a_3} g_{jk} g_{jl}, \qquad 1 \leqq k, l \leqq s.$$

Let S be an $s \times s$ symmetric matrix given by

$$S_{k,l} = h_{kl} + \delta_{kl} \|x_k\|_2^2.$$

Then

$$z^t S z = \sum_{j=1}^{s} (\|x_j\|_2 z_j)^2 + \sum_{j=1}^{a_3} \left(\sum_{k=1}^{s} g_{jk} z_k \right)^2, \qquad z \in \mathbb{R}^s,$$

and therefore the symmetric bilinear form defined by S is positive definite. We also note that S is diagonal if and only if $C = 0$.

THEOREM 4.3. *Let B be the matrix as above, and let A be a nonnegative matrix such that $AB = BA$ and* rank AB = rank A. *Let $A_0 = (\beta_{ij})$, $1 \leqq i, j \leqq s$, be the coefficient matrix of A with respect to B described in Lemma* 4.2.

Then $Ax = b$ has a nonnegative least squares solution for all nonnegative vectors $b \in R(B)$ if and only if $A_0 x = b$ has a nonnegative best approximate solution with respect to S for all nonnegative vectors $b \in \mathbb{R}^s$, where S is the symmetric bilinear form defined above.

Proof. As stated above, we assume

$$B = \begin{bmatrix} J & JD & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CJ & CJD & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where C, J, D are as in Lemma 4.1.

Let

$$u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_s \end{bmatrix} \in \mathbb{R}^s.$$

Define $u^\mu \in \mathbb{R}^{a_1}$, u^λ, $u^\nu \in \mathbb{R}^{a_1+a_2+a_3+a_4}$ by

$$(u^\lambda)_i = \begin{cases} u_l, & i = n_l + 1, \quad 1 \leqq l \leqq s, \\ 0 & \text{otherwise}, \end{cases}$$

i.e., $u^\lambda = [u_1 0 \cdots 0\ u_2 0 \cdots 0\ u_s 0 \cdots 0]^t$,

$$u^\mu = \begin{bmatrix} u_1 x_1 \\ u_2 x_2 \\ \vdots \\ u_s x_s \end{bmatrix}, \qquad u^\nu = \begin{bmatrix} u^\mu \\ 0_{a_2} \\ Cu^\mu \\ 0_{a_4} \end{bmatrix},$$

where x_i, $1 \leqq i \leqq s$, are the vectors appearing in the representation of the matrix B. We note that ν is an isomorphism from $\mathbb{R}^s$ onto $R(B)$, and indeed, ν maps nonnegative vectors in $\mathbb{R}^s$ to nonnegative vectors in $R(B)$.

Let

$$v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \end{bmatrix} \in \mathbb{R}^{a_1+a_2+a_3+a_4}.$$

We define $v^\phi \in \mathbb{R}^s$, $v^\psi \in \mathbb{R}^{a_1+a_2+a_3+a_4}$ as follows:

$$(v^\phi)_l = v_{n_l+1}, \quad 1 \leqq l \leqq s, \qquad v^\psi = ((X^{-1}SX^{-1})v^\phi)^\lambda,$$

where X is an $s \times s$ matrix such that $X_{k,l} = \delta_{kl} x_{1,k}$. We claim

$$(u^\nu)^\phi = Xu, \qquad u \in \mathbb{R}^s \tag{15}$$

$$((u^\nu)^\psi)^\phi = X^{-1}Su, \qquad u \in \mathbb{R}^s, \tag{16}$$

$$v - v^\psi \in R(B)^\perp, \qquad v \in R(B). \tag{17}$$

Since $(u^\nu)_{n_l+1} = x_{1,l} u_l$, $1 \leqq l \leqq s$, claim (15) follows immediately. Further, since $(v^\psi)^\phi = (X^{-1}SX^{-1})v^\phi$, claim (16) follows from claim (15). We now proceed to prove claim (17). Since $\{(e_l)^\nu | 1 \leqq l \leqq s\}$ is clearly a basis of $R(B)$, and since the operation ψ is linear, it suffices to prove the claim (17) for $v = (e_l)^\nu$. By definition of ν, we have

$$((e_l)^\nu)_i = \begin{cases} x_{j,l}, & i = n_l + j, \quad 1 \leqq j \leqq m_l, \\ g_{jl}, & i = a_1 + a_2 + j, \quad 1 \leqq j \leqq a_3, \\ 0 & \text{otherwise}. \end{cases}$$

By (16),

$$(((e_l)^\nu)^\psi)_i = \begin{cases} \dfrac{1}{x_{1,k}}(h_{kl} + \delta_{kl}\|x_k\|_2^2), & i = n_k + 1, \quad 1 \leqq k \leqq s \\ 0 & \text{otherwise.} \end{cases}$$

By actual computations we obtain

$$(((e_l)^\nu)^\psi)^t(e_k)^\nu = h_{kl} + \delta_{kl}\|x_k\|_2^2 = \sum_{j=1}^{a_3} g_{jl}g_{jk} + \delta_{kl}\sum_{i=1}^{m_k}(x_{i,k})^2 = ((e_l)^\nu)^t(e_k)^\nu.$$

Therefore

$$((e_l)^\nu - ((e_l)^\nu)^\psi)^t(e_k)^\nu = 0, \qquad 1 \leqq l, k \leqq s.$$

Hence

$$(e_l)^\nu - ((e_l)^\nu)^\psi \in R(B)^\perp, \qquad 1 \leqq l \leqq s.$$

This proves our claim (17).

Now assume $Ax = b$ has a nonnegative least squares solution for all nonnegative vectors $b \in R(B)$, and let c be an arbitrary nonnegative vector in $\mathbb{R}^s$. Since $c^\nu \in R(B)$, $Ax = c^\nu$ has a nonnegative least squares solution, say

$$f = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \end{bmatrix}.$$

Then

$$c^\nu = \sum_{i=1}^{a_1+a_2+a_3+a_4} f_i(A)^i + w_1, \qquad w_1 \in R(A)^\perp. \tag{18}$$

Further, since each $(A)^j$ can be expressed as a nonnegative linear combination of $(A)^{n_1+1}, \cdots, (A)^{n_s+1}$, we may assume in (18) that $f_i = 0$, $i \neq n_l + 1$, $1 \leqq l \leqq s$.

Next, we claim

$$(A)^{n_l+1} = y_{1,l}((A_0)^l)^\nu, \qquad 1 \leqq l \leqq s. \tag{19}$$

To prove (19) let

$$d_l = \begin{bmatrix} \beta_{1l}x_1 \\ \beta_{2l}x_2 \\ \vdots \\ \beta_{sl}x_s \end{bmatrix}.$$

Then

$$(A)^{n_l+1} = \begin{bmatrix} y_{1,l}d_l \\ 0 \\ Cy_{1,l}d_l \\ 0 \end{bmatrix}.$$

Therefore,

$$(A)^{n_l+1} = y_{1,l}((A_0)^l)^\nu, \qquad 1 \leqq l \leqq s,$$

as desired.

Then, by (18), (19) and by the assertion following (18), we have

$$c^{\nu} = \sum_{l=1}^{s} f_{n_l+1} y_{1,l} ((A_0)^l)^{\nu} + w_1. \tag{20}$$

Set

$$w_2 = c^{\nu} - (c^{\nu})^{\psi},$$
$$z_l = ((A_0)^l)^{\nu} - (((A_0)^l)^{\nu})^{\psi}, \qquad 1 \leqq l \leqq s.$$

Then by (17) $w_2, z_l \in R(B)^{\perp} \subseteq R(A)^{\perp}$. Also, by (20),

$$(c^{\nu})^{\psi} = \sum_{l=1}^{s} f_{n_l+1} y_{1,l} (((A_0)^l)^{\nu})^{\psi} + w_3, \tag{21}$$

where

$$w_3 = w_1 + \left(\sum_{l=1}^{s} f_{n_l+1} y_{1,l} z_l \right) - w_2,$$

and thus $w_3 \in R(A)^{\perp}$. Set $w' = (X^{-1}S)^{-1}(w_3)^{\phi}$. By (16) and (21),

$$(X^{-1}S)c = \sum_{l=1}^{s} f_{n_l+1} y_{1,l} ((X^{-1}S)(A_0)^l) + (X^{-1}S)w'.$$

It then follows that

$$c = \sum_{l=1}^{s} f_{n_l+1} y_{1,l} (A_0)^l + w'. \tag{22}$$

Since $w_3 \in R(A)^{\perp}$, we get from (19),

$$(((A_0)^l)^{\nu})^t w_3 = 0, \qquad 1 \leqq l \leqq s. \tag{23}$$

Also, by (21), $(w_3)_i = 0$, $i \neq n_l + 1$, $1 \leqq l \leqq s$. Therefore, we may rewrite (23) as

$$(X(A_0)^l)^t (w_3)^{\phi} = 0 \tag{23'}$$

by using (15). Then by (23'), together with the definition of w', we have

$$((A_0)^l)^t S w' = 0.$$

Hence, $w' \in R(A_0)^{\perp_S}$. Thus (22) gives us a nonnegative best approximate solution of $A_0 x = c$ with respect to the norm S.

We can retrace the steps back to prove the "if" part of the theorem, completing the proof.

Combining Theorems 3.7 and 4.3, we obtain the following main result.

THEOREM 4.4. *Let B be a nonnegative idempotent matrix. Let A be a nonnegative matrix such that $AB = BA$ and* rank (AB) = rank A. *Let A_0 be a coefficient matrix of A with respect to B. Suppose that the equation $Ax = b$ has a nonnegative least squares solution for all nonnegative vectors $b \in R(B)$. Then there exist permutation matrices P, Q such that A_0 can be expressed in the form*

$$PA_0Q = \begin{bmatrix} G & GL \\ 0 & 0 \end{bmatrix},$$

where

$$G = \begin{bmatrix} z_1 & & & \\ & z_2 & & \\ & & \ddots & \\ & & & z_k \end{bmatrix},$$

z_i are positive vectors and L is a nonnegative matrix, or equivalently, A_0 has a nonnegative $\{1, 3\}$-inverse.

We give an example to demonstrate that the converse of Theorem 4.4 is not necessarily true.

Example 4.5. Let

$$B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 2 & 1 & 1 & 0 \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 3 & 3 & 1 & 0 \end{bmatrix}.$$

Then $B = B^2$, $AB = BA$, rank (AB) = rank $A = 2$ and a coefficient matrix A_0 of A is given by

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Although A_0 has a nonnegative $\{1, 3\}$-inverse, we may verify that the system $Ax = b$, where

$$b = \begin{bmatrix} 0 \\ 13 \\ 0 \\ 13 \end{bmatrix} \in R(B),$$

does not possess a nonnegative least squares solution. For, if we write $b = b_1 + b$, where

$$b_1 = \begin{bmatrix} 5 \\ 5 \\ -1 \\ 14 \end{bmatrix} \in R(A), \qquad b_2 = \begin{bmatrix} -5 \\ 8 \\ 1 \\ -1 \end{bmatrix} \in R(A)^{\perp},$$

then by Sublemma 3.4 a least squares solution x_0 must satisfy $Ax_0 = b_1$. But then x_0 cannot be nonnegative.

Remarks 4.6. (1) Recall from Remark 3.9 that if the positive definite symmetric bilinear form S is diagonal, then the existence of a nonnegative $\{1, 3\}$-inverse of a matrix A is equivalent to the existence of nonnegative best approximate solution of $Ax = b$ for all nonnegative vectors b. Also recall that the symmetric bilinear form S in Theorem 4.3 is diagonal if and only if the matrix C in Lemma 4.1 is zero. Therefore, it follows that the converse of Theorem 4.4 holds if $C = 0$.

(2) Example 4.5 shows that the converse of Theorem 4.4 does not hold. Nevertheless, we can show that if an $m \times m$ matrix A is as in the conclusion of Theorem 4.4, then there exists a positive definite symmetric bilinear form S over $\mathbb{R}^m$ such that

$Ax = b$ has a nonnegative best approximate solution with respect to S for all nonnegative vectors $b \in R(B)$.

(3) Let B be an $m \times n$ (not necessarily square) nonnegative matrix of rank s such that

$$(P_0BQ_0)_{i,i} \neq 0, \quad 1 \leqq i \leqq s, \qquad (P_0BQ_0)_{i,j} = 0, \quad 1 \leqq i, j \leqq s, \quad i \neq j$$

for suitable permutation matrices P_0 and Q_0. Let A be an $m \times l$ nonnegative matrix such that $R(A) \subseteq R(B)$. Let A_0' be the matrix consisting of the first s rows of P_0AQ_0. With A, B and A_0' as above, arguments similar to the proof of Theorem 4.3 prove the following:

If $Ax = b$ has a nonnegative least squares solution for all nonnegative $b \in R(B)$, then A_0' has a nonnegative $\{1, 3\}$-inverse.

In case A and B are as in Theorem 4.4, we give below the relation between a coefficient matrix $A_0 = (\beta_{ij})$ and the matrix A_0'. It follows as a consequence that the existence of a nonnegative $\{1, 3\}$-inverse of A_0' implies that of A_0 and vice versa. Let P be as in Lemma 4.1. Also let the notation be as in the proof of Lemma 4.2 with B replaced by PBP^t. Further, let P_1 be a permutation matrix which sends the $(n_i + 1)$th row to the ith row. Set $P_0 = P_1P$ and $Q_0 = P_0^t$. Then P_0BQ_0 is in the form stated at the beginning of this remark. With this choice of P_0 and Q_0, we have

$$\beta_{ij} = \frac{(A_0')_{i,j}}{x_{1,i}y_{1,j}}, \qquad 1 \leqq i, j \leqq s.$$

Since every column of A_0' can be written as a nonnegative linear combination of the first s columns of A_0', the existence of a nonnegative $\{1, 3\}$-inverse of one of A_0 or A_0' implies that of the other.

Acknowledgment. The authors express their sincere thanks to Professor D. K. Ray-Chaudhuri, Chairman, Department of Mathematics, Ohio State University, for providing them the opportunity of working together on this paper.

REFERENCES

[1] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[2] ———, *Monotonicity and the generalized inverse*, SIAM J. Appl. Math., 22 (1972), pp. 155–161.

[3] ———, *Inverses of nonnegative matrices*, Linear and Multilinear Algebra, 2 (1974), pp. 161–172.

[4] P. FLOR, *On groups of nonnegative matrices*, Compositio Math., 21 (1969), pp. 376–382.

[5] S. FRIEDLAND AND H. SCHNEIDER, *The growth of powers of a nonnegative matrix*, this Journal, 1 (1980), pp. 185–200.

[6] F. J. HALL AND I. J. KATZ, *Nonnegative generalized inverses*, Linear Algebra Appl., to appear.

[7] E. HAYNSWORTH AND J. R. WALL, *Group inverses of certain nonnegative matrices*, Linear Algebra Appl., 25 (1979), pp. 271–288.

[8] S. K. JAIN, V. K. GOEL AND EDWARD KWAK, *Nonnegative matrices having same nonnegative Moore-Penrose and group inverses*, Linear and Multilinear Algebra, 7 (1979), pp. 59–72.

[9] ———, *Decomposition of nonnegative group-monotone matrices*, Trans. Amer. Math. Soc., 257 (1980), pp. 371–385.

[10] S. K. JAIN AND L. E. SNYDER, *Nonnegative λ-monotone matrices*, this Journal, 2 (1981), pp. 66–76.

[11] R. J. PLEMMONS AND R. E. CLINE, *The generalized inverse of a nonnegative matrix*, Proc. Amer. HMath. Soc., 31 (1972), pp. 46–50.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 2, June 1982

0196-5212/82/0302-0008 $01.00/0

DECOMPOSITION OF DIRECTED GRAPHS*

WILLIAM H. CUNNINGHAM†

Abstract. A composition for directed graphs which generalizes the substitution (or X-join) composition of graphs and digraphs, as well as the graph version of set-family composition, is described. It is proved that a general decomposition theory can be applied to the resulting digraph decomposition. A consequence is a theorem which asserts the uniqueness of a decomposition of any digraph, each member of the decomposition being either indecomposable or "special". The special digraphs are completely characterized; they are members of a few interesting classes. Efficient decomposition algorithms are also presented.

1. Introduction. Throughout this paper *digraph* or *directed graph* means "simple finite directed graph"; that is, the vertex-set is a finite set $V(G)$ and the edge-set $E(G)$ is a subset of $\{(u, v): u, v \in V(G), u \neq v\}$. For the most part, our terminology follows Bondy and Murty [1]. Let G_1, G_2 be directed graphs having vertex-sets $V_1 \cup \{v\}$, $V_2 \cup \{v\}$ respectively, where $\{V_1, V_2\}$ is a partition of V and $v \notin V$. We define a digraph $G = G_1 * G_2$, the *composition* of G_1 with G_2, to have vertex-set V and edge-set $\{(x, y): (x, y) \in E(G_1) \cup E(G_2), x \neq v \neq y\} \cup \{(x, y): (x, v) \in E(G_1)$ and $(v, y) \in E(G_2)$ or $(x, v) \in E(G_2)$ and $(v, y) \in E(G_1)\}$. This composition is illustrated in Fig. 1. This paper presents a unique decomposition theory for this digraph composition based on a general decomposition theory [6] and also efficient decomposition algorithms.

We begin by describing some interesting special cases of the composition. Clearly, G is symmetric (satisfies $(x, y) \in E(G)$ if and only if $(y, x) \in E(G)$ for all $x, y \in V(G)$) if and only if G_1 and G_2 are, so one special case is a composition for undirected graphs. Equivalently, this is a composition for families of sets (hypergraphs) each of whose members has cardinality exactly 2. In this latter context, the undirected-graph composition is also a special case of a set-family composition investigated previously [6, § 5]. We remark that this graph composition is powerful enough to encompass the standard notion of graph separability. Specifically, with a single trivial exception, any connected graph which is separable (has a cut vertex) can be expressed as a $*$-composition of smaller graphs.

Another important special case occurs in a less transparent way. Suppose that H, J are digraphs having disjoint vertex-sets and let $v \in V(H)$. Then $H[J; v]$, the *substitution composition* or "X-join" of H with J, is the digraph having vertex-set $(V(H) \cup V(J)) \backslash \{v\}$ and edge-set $\{(x, y): (x, y) \in E(H), x \neq v \neq y\} \cup E(J) \cup \{(x, y): (x, v) \in E(H), y \in V(J)$ or $(v, y) \in E(H), x \in V(J)\}$. Let us denote by vJ the

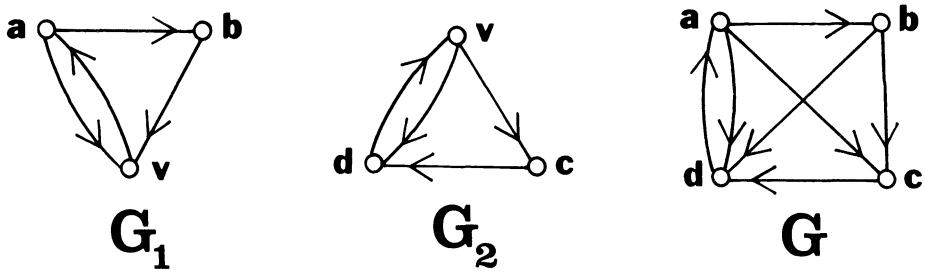


FIG. 1

* Received by the editors February 5, 1981, and in final revised form October 19, 1981. This research was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

† Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6. Currently on leave at Institut für Operations Research, Universität Bonn, with research supported in part by SFB 21(DFG).

digraph having vertex-set $V(J)\cup\{v\}$ and edge-set $E(J)\cup\{(x, v): x\in V(J)\}\cup\{(v, x): x\in V(J)\}$. (A digraph such as vG is said to be *pointed* at v.) It is easy to see that $H[J; v]=H*(vJ)$. We can obtain as a further special case the substitution [3] or X-join [11] composition for undirected graphs, either by considering symmetric digraphs in the digraph substitution, or by applying the pointing device to the undirected-graph composition above. Thus, the composition $*$ for digraphs generalizes the ordinary graph substitution in two different directions. Other attractive special classes of digraphs to which the digraph substitution can be applied include acyclic digraphs or their transitive closures (partial orders).

There are a number of combinatorial optimization problems which can be solved (more efficiently) on the composition graph by solving similar problem instances on the smaller graphs. Examples of such problems are sequencing problems on acyclic digraphs [12] and the optimal stable set problem for undirected graphs [3]. Both of these applications concern compositions of the substitution type. Here we describe a similar result for the more general composition, which generalizes Chvátal's result. Where G is an undirected graph, a set $S\subseteq V(G)$ is *stable* if no two vertices in S are adjacent in G. For a given real-valued weight vector $(c_u: u\in V(G))$, the *optimal stable set problem* is to maximize $\sum(c_u: u\in S)$ over stable sets S of G. Suppose that $G=G_1*G_2$, where G_1 and G_2 have the common vertex v. Let S_1 be an optimal stable set in G_1-v, and let S_1' be an optimal stable set in $G_1-(\{v\}\cup N(v))$, where $N(v)$ is the set of vertices adjacent to v in G_1. Let S_2 be an optimal stable set in G_2, where c_v is defined to be $\sum(c_u: u\in S_1)-\sum(c_u: u\in S_1')$. Then S is an optimal stable set in G, where $S=S_2\cup S_1'$ if $v\notin S_2$ and $S=(S_2\cup S_1)\backslash\{v\}$ otherwise. This observation applies also to "iterated" decompositions. Because there exists an efficient algorithm to find an expression for a given graph as a composition of smaller graphs (if possible), these techniques can be used to enlarge the class of graphs for which the optimal stable set problem can be solved in polynomial time. (As an example of an optimization problem in which the most general digraph composition is useful, we mention the optimum dominating set problem. A set $D\subseteq V(G)$ is *dominating* if, for every $v\in V(G)\backslash D$, there exists $u\in D$ with $(u, v)\in E(G)$.)

In the next section we describe terminology from decomposition theory [6] and state the main unique decomposition theorems. The first such theorem asserts the uniqueness of a decomposition of a diconnected digraph into indecomposable and certain highly decomposable digraphs. This result is improved by the complete characterization of the highly decomposable digraphs. The rest of the section is concerned with the application of these results to the special classes mentioned above. Section 3 presents the proofs of these uniqueness theorems. Two main steps are needed. The first, which shows that the properties required to use the theory of [6] are satisfied, is essentially easy. The second, proving the characterization of the highly decomposable digraphs, is more difficult. The last section presents polynomial-time algorithms for carrying out all decompositions discussed in the paper.

2. Unique decomposition theorems. If $G=G_1*G_2$ as at the beginning of § 1 and, in addition $|V_1|\geqq 2\leqq|V_2|$, we say that $\{G_1, G_2\}$ is a *simple decomposition* of G and write $G\to\{G_1, G_2\}$. We call $\{V_1, V_2\}$ the *split* of G associated with the simple decomposition and v the associated *marker* element. A (general) *decomposition* of a digraph G is defined inductively to be either $\{G\}$ or a set D' of digraphs obtained from a decomposition D of G by replacing a member G_1 of D by the members of a simple decomposition of G_1, where the marker of this simple decomposition is not a vertex of any member of D. If D'' is obtained from D by a (nonempty) sequence of

operations of the kind described above, then D'' is said to be a (*strict*) *refinement* of D. If the sequence consists of exactly one operation, the refinement is *simple*.

We can associate a graph T with any decomposition D of a digraph G. The vertices of T are the members of D and edges are the markers of D; each marker joins in T the two members of D of which it is a vertex. It is clear that T is a tree. This "decomposition tree" provides a useful way to visualize a decomposition.

Two decompositions D, D' of G are *equivalent* if D' can be obtained from D by replacing some of the markers of D by markers of D'. All unique decomposition theorems of this paper involve uniqueness "up to equivalence", but we will not include this phrase in their statements. The decomposition D of G is *minimal* with some property P if D has P and there does not exist a decomposition D' of G also having P such that D is a strict refinement of D'. A decomposition D is *trivial* if $|D| = 1$. A digraph D is *prime* if it has no nontrivial decomposition.

If G is a digraph and $A \subseteq V(G)$, we denote by $\delta(A)$ the set $\{(x, y): (x, y) \in E(G), x \in A, y \notin A\}$. A digraph G is *diconnected* (or strongly connected) if for all A, $\phi \subset A \subset V(G)$ we have $\delta(A) \neq \phi$. We state without proof some easy but useful results. (The symbol δ_1 in Proposition 2 has the expected meaning.)

PROPOSITION 1. *$\{V_1, V_2\}$ is a split of the digraph G if and only if, for $i = 1$ and 2, $(a, b), (c, d) \in \delta(V_i)$ implies $(a, d) \in \delta(V_i)$.*

PROPOSITION 2. *If $G \to \{G_1, G_2\}$ with associated split $\{V_1, V_2\}$ and marker v and $A \subseteq V_1$, then*

(a) *$\delta_1(A) = \phi$ if and only if $\delta(A) = \phi$;*

(b) *$\delta_1(A \cup \{v\}) = \phi$ if and only if $\delta(A \cup V_1) = \phi$.*

COROLLARY 1. *If $G \to \{G_1, G_2\}$, then G is diconnected if and only if G_1, G_2 are diconnected.*

Because of Proposition 2 and Corollary 1, we are justified in restricting the digraph decomposition theory to diconnected digraphs. We know that this class is closed under composition and decomposition by Corollary 1. But Proposition 2 tells us even more; roughly speaking, it says that G lacks diconnectivity in the same ways that G_1 and G_2 do.

It is, perhaps, natural to hope that each diconnected digraph G would have a unique decomposition consisting of prime digraphs, but this is not the case. Consider, for example, the *dicomplete* digraphs: $E(G)$ consists of every ordered pair of distinct elements of $V(G)$. A dicomplete digraph having four or more vertices has inequivalent prime decompositions; any dicomplete digraph having six or more vertices has prime decompositions having nonisomorphic decomposition trees. In fact, dicomplete digraphs are examples of "brittle" digraphs. A digraph G is *brittle* if $|V(G)| \geqq 4$ and every partition $\{V_1, V_2\}$ of $V(G)$ satisfying $|V_1| \geqq 2 \leqq |V_2|$ is a split of G. The brittle digraphs comprise the first of two classes of highly decomposable digraphs which play a special role in the decomposition theory. A digraph G is *semibrittle* if $|V(G)| \geqq 4$ and there exists an ordering $v_0, v_1, \cdots, v_{n-1}$ of $V(G)$ such that the splits of G are precisely the partitions $\{\{v_i, v_{i+1}, \cdots, v_{i+j-1}\}, \{v_{i+j}, \cdots, v_{i-1}\}\}$, where $0 \leqq i \leqq n-1$, $2 \leqq j \leqq n-2$ and subscripts are taken modulo n. An example of a semibrittle digraph is a digraph consisting of a directed cycle of length at least 4. The digraph G of Fig. 1 is also semibrittle, and this particular example provides an even stronger illustration of the lack of uniqueness of prime decompositions. The digraphs H_1, H_2 of Fig. 2 comprise a simple decomposition of G, and the members of $\{H_1, H_2\}$ are not even pairwise isomorphic to the members of the decomposition $\{G_1, G_2\}$ of Fig. 1.

In the next section, we apply the theory of [6] to prove the following unique decomposition result.

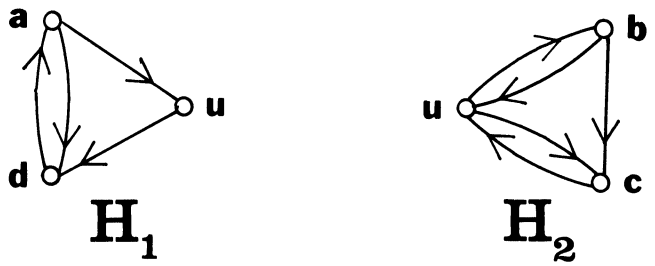


FIG. 2

THEOREM 1. *Each diconnected digraph has a unique minimal decomposition, each of whose members is prime, brittle or semibrittle.*

The main uniqueness result for digraph decomposition is obtained from Theorem 1 by characterizing the brittle and semibrittle digraphs. A *distar* is a digraph G such that $E(G)=\{(u, v): v \in V(G)\backslash\{u\}\}\cup\{(v, u): v \in V(G)\backslash\{u\}\}$ for some $u \in V(G)$. (The vertex u is called the *center* of the star.) It is not hard to see that distars are diconnected and brittle. A *transitive tournament* is a digraph G such that for some ordering v_1, $v_2, \cdots, v_{n-1}$ of $V(G)$ we have $E(G)=\{(v_i, v_j): 1 \leqq i<j \leqq n-1\}$. Clearly, a transitive tournament is not diconnected (unless $n=2$); however, some interesting diconnected digraphs are constructed from transitive tournaments. A *circle of transitive tournaments* (CTT) is a digraph obtained from a sequence $T_1, T_2, \cdots, T_k$ of transitive tournaments, where $|V(T_i)| \geqq 2$ for each i, by identifying the last vertex of T_i with the first vertex of T_{i+1} for $1 \leqq i \leqq k-1$ and identifying the last vertex of T_k with the first vertex of T_1. More formally, G is a CTT if, for some ordering $v_0, v_1, \cdots, v_{n-1}$ of $V(G)$ and integers $0=p_1<p_2<\cdots<p_k<p_{k+1}=n$, we have $E(G)=\{(v_i, v_j): p_l \leqq i<j \leqq p_{l+1}$ for some $l, 1 \leqq l \leqq k\}$, where v_n means v_0. (Note, however, that $(v_0, v_0) \notin E(G)$.) The vertices $v_{p_1}, v_{p_2}, \cdots, v_{p_k}$ are called the *hinges* of the CTT. We say that the CTT is of *type* $(p_2-p_1, p_3-p_2, \cdots, p_{k+1}-p_k)$. There is a CTT on n vertices for each ordered partition of the integer n into positive integers, up to a cyclic permutation of the partition. The seven distinct (up to isomorphism) CTT's on 5 vertices are illustrated in Fig. 3.

The CTT's are diconnected and semibrittle. Their appearance enriches the digraph decomposition theory considerably. First, the CTT's for $k>1$ are new; they do not appear in the more restricted theories which we have mentioned as special cases of the present theory. Second, they are numerous; up to isomorphism there is just one dicomplete and one distar having n vertices, but the number of CTT's grows exponentially with n. We can now state the main unique decomposition theorem for digraphs.

THEOREM 2. *Each diconnected digraph has a unique minimal decomposition, each of whose members is prime, dicomplete, a distar or a circle of transitive tournaments.*

Theorems 1 and 2 will be proved in § 3. We devote the remainder of this section to an investigation of their applications to the less general situations mentioned earlier. A decomposition theorem for undirected graphs is obtained by restricting attention to symmetric digraphs in Theorem 2. (We observe that $G_1 * G_2$ is symmetric if and only if G_1 and G_2 are symmetric.) The assumption of diconnectivity for digraphs reduces to connectivity for undirected graphs. Dicomplete digraphs and distars are symmetric, but no CTT is symmetric. *Complete* graphs and *stars* are undirected graphs defined as expected. The resulting theorem first appeared in [5, (625)] and is a special case of [6, Thm. 11].

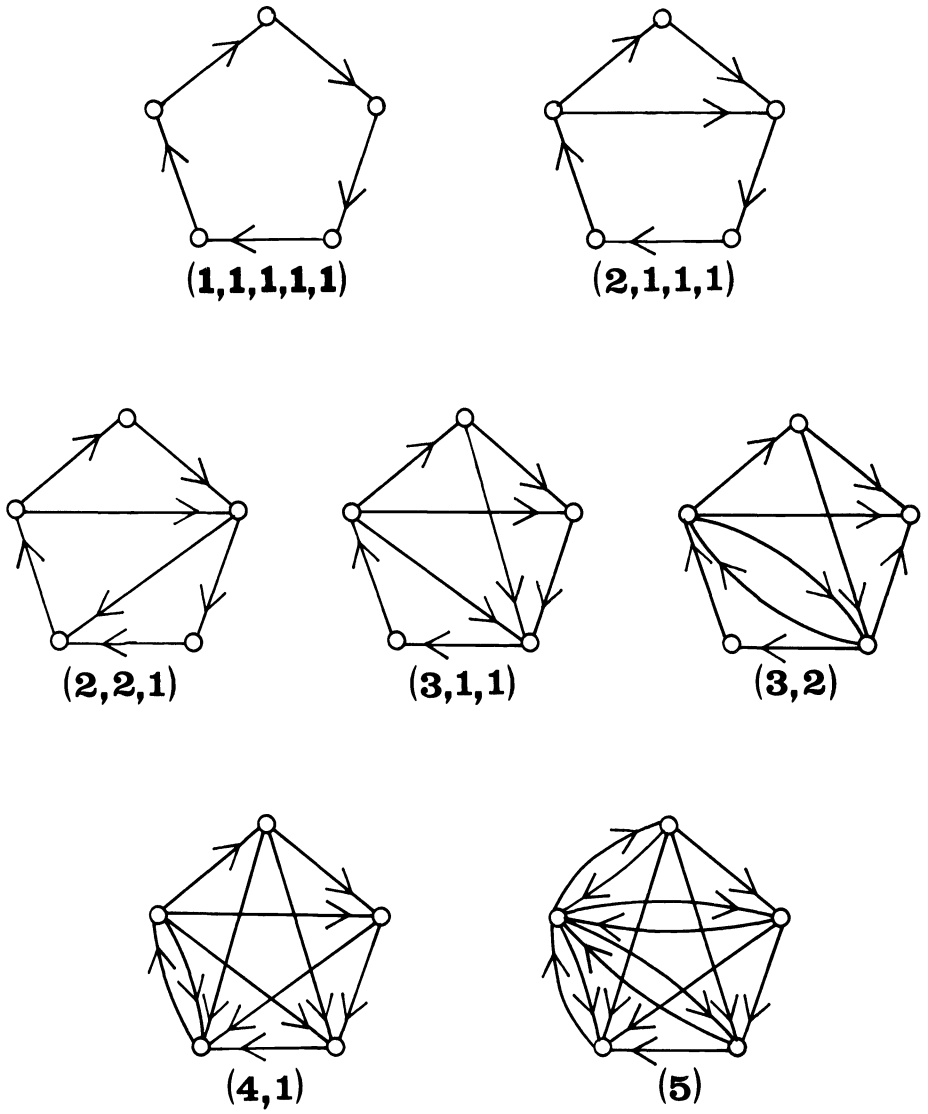


FIG. 3

THEOREM 3. *Each connected graph has a unique minimal decomposition, each of whose members is prime, complete or a star.*

Now let us consider a decomposition theory for the substitution-type digraph composition. Since a substitution composition is not determined by the two digraphs G_1, G_2 being composed and their order (the special vertex of G_1 is also needed), it is convenient (as in [6]) to consider the objects being decomposed as pairs (G, v), where v is an element not in $V(G)$. We say that $\{(G_1, v), (G_2, w)\}$ is a *simple factorization* of (G, v) if $G = G_1[G_2; w]$. The vertex w is called the *marker* associated with the factorization. A (general) *factorization* of (G, v) is defined inductively to be either $\{(G, v)\}$ or a set D' obtained from a factorization D of (G, v) by replacing a member (G_1, v_1) of D by the members of a simple factorization of (G_1, v_1) such that

the marker of this simple factorization is neither a vertex of G nor an element v' such that $(G', v') \in D$ for some G'. We define (*simple*) *refinement*, *marker* (of a factorization), *equivalent*, *trivial* and *minimal* just as for decompositions. A *component* of a factorization D is a digraph H such that $(H, w) \in D$ for some w. A digraph G is *irreducible* if (G, v) has no nontrivial factorization.

Given a factorization D of (G, v) we can form a directed graph T as follows. The vertices of T are the components of D and the edges are its markers. The marker w is directed from G_1 to G_2, where G_1 is the component of D such that $(G_1, w) \in D$ and G_2 is the component of D of which w is a vertex. It is easy to see that T is a tree having the property that every vertex but one has exactly one edge directed away from it; the exceptional vertex is the component G' of D such that $(G', v) \in D$. This component G' is sometimes called the *quotient* of D.

Recall the definition of the digraph vJ from § 1 and the fact that $H[J; v] = H * (vJ)$. An equivalent statement is that $\{(H, w), (J, v)\}$ is a simple factorization of (G, w) if and only if $\{H, vJ\}$ is a simple decomposition of G. This result is easily generalized to arbitrary decompositions and factorizations by making use of their inductive definitions.

PROPOSITION 3. *Let G be a digraph and $v \notin V(G)$. Then D is a factorization of (G, v) if and only if $D' = \{v'G' : (G', v') \in D\}$ is a decomposition of vG.*

We can observe that vG is diconnected for any digraph G, so Proposition 3 allows us to obtain a theory for the substitution decomposition which applies to all digraphs. Clearly vG is prime if and only if G is irreducible. It is also easy to see which digraphs G have the property that vG is dicomplete, a distar or a CTT. They are dicomplete digraphs, edgeless digraphs and transitive tournaments, arising, respectively, from dicomplete digraphs, distars and CTT's of type $(|V(G)|)$. The resulting unique decomposition theorem for digraph substitution can now be stated.

THEOREM 4. *Let G be a digraph and $v \notin V(G)$. Then (G, v) has a unique minimal factorization, each of whose components is irreducible, dicomplete, edgeless or a transitive tournament.*

Theorems 3 and 4 are specializations of Theorem 2. On the other hand the next result, on substitution decomposition of undirected graphs, is a specialization of both Theorem 3 and Theorem 4. We can obtain it from Theorem 4 by considering only symmetric digraphs, or we can obtain it from Theorem 3 in the same way that Theorem 4 was obtained from Theorem 2. This result appears in [5, (733)] and in [6, § 7].

THEOREM 5. *Let G be a graph and let $v \notin V(G)$. Then (G, v) has a unique minimal factorization, each of whose components is irreducible, complete or edgeless.*

Another special class to which the substitution decomposition theory can be applied is the class of acyclic digraphs, which is closed under substitution composition and decomposition. Using Theorem 4 and the fact that dicomplete digraphs are not acyclic, we obtain the following result.

THEOREM 6. *Let G be an acyclic digraph and let $v \notin V(G)$. Then (G, v) has a unique minimal factorization, each of whose components is an acyclic digraph which is irreducible, edgeless or a transitive tournament.*

One can also consider the class of transitive acyclic digraphs. Theorem 6 remains true with "acyclic" replaced by "transitive acyclic". It is interesting to notice the connection with partially ordered sets, which are equivalent to transitive acyclic digraphs. Of course, the sets $C \subseteq V(G)$ such that $\{C, (V(G) \backslash C) \cup \{v\}\}$ is a split of vG play an important role in the theory. These are called "job modules" in [12], where the partial order is that imposed by a precedence relation among jobs in a sequencing problem. Moreover, the partially ordered sets associated with transitive tournaments

and edgeless digraphs are, appropriately, *chains* (totally ordered sets) and *antichains* (totally unordered sets). (The familiar class of *series-parallel* posets consists of posets having a factorization whose components are chains and antichains.)

Finally, we remark that a weaker type of unique decomposition theorem can be proved for prime decompositions in many of the applications. The idea is that any prime decomposition D of a digraph G is a refinement of the "standard" decomposition D' whose uniqueness is asserted in Theorem 2. Therefore, D is obtained by replacing each brittle or semibrittle member of D' by a prime decomposition of that member. Dicomplete digraphs, distars and CTT's of type (n) all have the property that any decomposition of one of them consists of digraphs of the same type. That is, any decomposition of a distar consists of distars and so on. Thus, for example, any prime decomposition of a distar having at least three vertices consists of distars having three vertices. We can apply this observation to any of Theorems 3, 4, 5, 6 to obtain theorems on the uniqueness *up to isomorphism* of prime decompositions; as an illustration the resulting corollary to Theorem 4 is stated below as Theorem 7. However, the same approach cannot be applied to the most general uniqueness result, Theorem 2. This can be seen from the example of Figs. 1 and 2.

THEOREM 7. *Let G be a digraph and $v \notin V(G)$. The components of any factorization of (G, v), each of whose components is irreducible, are unique up to isomorphism.*

We observe that the same argument allows us to conclude that the quotient in Theorem 7 is unique up to isomorphism. It follows that all factorizations of (G, v) having irreducible quotients, have isomorphic quotients. This result, for a restricted class of factorizations is due to Maurer [9], [10].

3. Proofs. In order to prove Theorem 1, we use the general theory developed in [6]. Let $\mathcal{G}$ be the class of diconnected digraphs. For each $G \in \mathcal{G}$, $V(G)$ (of course) denotes the vertex-set of G, and $\rightarrow$, as defined in § 2, is a relation associating elements G of $\mathcal{G}$ with two-element subsets $\{G_1, G_2\}$ of $\mathcal{G}$. A triple such as $(\mathcal{G}, V, \rightarrow)$ is defined in [6] to be a *decomposition frame* if four axioms are satisfied. These are F1–F4 of Theorem 8 below; in other words, the content of Theorem 8 is that $(\mathcal{G}, V, \rightarrow)$ is a decomposition frame.

THEOREM 8. $(\mathcal{G}, V, \rightarrow)$ *satisfies* F1–F4:

F1. *If $G \in \mathcal{G}$ and $G \rightarrow \{G_1, G_2\}$, then for some $v \notin V(G)$ and some partition $\{V_1, V_2\}$ of $V(G)$ with $|V_1| \geqq 2 \leqq |V_2|$, we have $V(G_1) = V_1 \cup \{v\}$, $V(G_2) = V_2 \cup \{v\}$.*

F2. *For a split $\{V_1, V_2\}$ of $G \in \mathcal{G}$ and $v \notin V(G)$, there is exactly one simple decomposition $\{G_1, G_2\}$ of G with marker v corresponding to $\{V_1, V_2\}$. (In the situation described in* F2, *we denote by $G(V_i; v)$ the digraph G_i, $i = 1$ and 2.)*

F3. *Let $\{V_1, V_2\}$ be a split of $G \in \mathcal{G}$, let $A \subset V_1$ and $v \notin V(G)$. Then $\{A, V(G) \backslash A\}$ is a split of G if and only if $\{A, (V_1 \cup \{v\}) \backslash A\}$ is a split of $G(V_1; v)$.*

F4. *Let $\{V_1, V_2\}, \{V_3, V_4\}$ be splits of $G \in \mathcal{G}$ such that $V_3 \subset V_1$ and let $v, w \notin V(G)$, $v \neq w$. Then $G(V_1; v)(V_3; w) = G(V_3; w)$, and $G(V_1; v)((V_1 \backslash V_3) \cup \{v\}; w) = G(V_4; w)((V_4 \backslash V_2) \cup \{w\}; v)$.*

Proof. Clearly, F1 holds. For F2, it is easy to see that we must have $E(G(V_1; v)) = \{(x, y): x, y \in V_1, (x, y) \in E(G)\} \cup \{(x, v): x \in V_1,$ there exists $y \in V_2$ with $(x, y) \in E(G)\} \cup \{(v, y): y \in V_1,$ there exists $x \in V_2$ with $(x, y) \in E(G)\}$, and similarly for $G(V_2; v)$.

Now consider F3. We abbreviate $G(V_i; v)$ to G_i for $i = 1$ and 2. Suppose that $\{A, (V_1 \backslash A) \cup \{v\}\}$ is a split of G_1 and let (a, b), $(c, d) \in \delta(A)$; we must show that $(a, d) \in E(G)$. Either (a, b) or (a, v) is an edge of G_1 and either (c, d) or (c, v) is an edge of G_1. Thus either (a, d) or (a, v) is an edge of G_1.

In the former case (a, d) is an edge of G, and in the latter case (v, d) is an edge of G_2, so (a, d) is an edge of G. The proof for the case (a, b), $(c, d) \in \delta(V(G)\backslash A)$ is similar. Now suppose that $\{A, V(G)\backslash A\}$ is a split of G with $A \subset V_1$; let (a, b), $(c, d) \in \delta_1(A)$. Then $(a, b) \in E(G)$, or $b = v$ and $(a, b') \in E(G)$ for some $b' \in V_2$. Similarly, $(c, d) \in E(G)$, or $d = v$ and $(c, d') \in E(G)$ for some $d' \in V_2$. Thus $(a, d) \in E(G)$, or $d = v$ and $(a, d') \in E(G)$ for some $d' \in V_2$. In either case $(a, d) \in E(G_1)$, as required. The proof for the case (a, b), $(c, d) \in \delta_1((V(G)\backslash A) \cup \{v\})$ is similar.

Finally, we consider F4. Let $G_1 = G(V_1; v)$ and let $G_3 = G(V_3; w)$. $E(G_1(V_3; w)) = \{(x, y) \in E(G_1)\colon x, y \in V_3\} \cup \{(x, w)\colon x \in V_3$ and, for some $y \in (V_1\backslash V_3) \cup \{v\}$, $(x, y) \in E(G_1)\} \cup \{(w, y)\colon y \in V_3$ and, for some $x \in (V_1\backslash V_3) \cup \{v\}$, $(x, y) \in E(G_1)\} = E(G_3)$. Now let $H = G(V_1; v)((V_1\backslash V_3) \cup \{v\}; w)$, and let $J = G(V_4; w)(V_4\backslash V_2) \cup \{w\}; v)$. Then $(v, w) \in E(H)$ if and only if there exists $(x, y) \in E(G)$ with $x \in V_2$ and $y \in V_3$, and this is also the condition for (v, w) to be an element of $E(J)$. Similarly, $(w, v) \in E(H)$ if and only if $(w, v) \in E(J)$. Finally, it is straightforward to check that the set of remaining edges of H or J is $\{(x, y) \in E(G)\colon x, y \in V_1\backslash V_3\} \cup \{(x, v)\colon x \in V_1\backslash V_3$ and, for some $y \in V_2$, $(x, y) \in E(G)\} \cup \{(v, y)\colon y \in V_1\backslash V_3$ and, for some $x \in V_2$, $(x, y) \in E(G)\} \cup \{(x, w)\colon x \in V_1\backslash V_3$ and, for some $y \in V_3$, $(x, y) \in E(G)\} \cup \{w, y)\colon y \in V_1\backslash V_3$ and, for some $x \in V_3$, $(x, y) \in E(G)\}$. Thus $H = J$, and the proof is complete. □

The next result shows that the decomposition frame $(\mathcal{G}, V, \rightarrow)$ has the *intersection property*; this property plays an important role in the theory presented in [6]. It is interesting to note that this result is not generally true for digraphs which are not diconnected.

THEOREM 9. *Let* $\{V_1, V_2\}$, $\{V_3, V_4\}$ *be splits of* $G \in \mathcal{G}$ *such that* $|V_1 \cap V_3| \geq 2$ *and* $V_1 \cup V_3 \neq V(G)$. *Then* $\{V_1 \cap V_3, V_2 \cup V_4\}$ *is a split of* G.

Proof. Let (a, b), $(c, d) \in \delta(V_1 \cap V_3)$. If (a, b), $(c, d) \in \delta(V_1)$, or if (a, b), $(c, d) \in \delta(V_2)$, we are done, so we may assume that $b \in V_1 \cap V_4$, $d \in V_2 \cap V_3$. Since G is diconnected, there exists an edge $(p, q) \in \delta(V_1 \cup V_3)$. In fact, we can choose (p, q) so that $p \in V_1 \cap V_3$. For suppose that $p \in V_1\backslash V_3$. (The case $p \in V_3\backslash V_1$ is similar.) Then (p, q), $(c, d) \in \delta(V_1)$, so $(c, q) \in E(G)$, and we can replace p by c. Now (p, q), $(a, b) \in \delta(V_3)$, so $(a, q) \in E(G)$. Finally, (a, q), $(c, d) \in \delta(V_1)$, so $(a, d) \in E(G)$, as required. A similar argument can be repeated for pairs (a, b), $(c, d) \in \delta(V_2 \cup V_4)$. □

[6, Thm. 4] states that any "object" of a decomposition frame having the intersection property has a unique minimal decomposition consisting of prime, brittle and semibrittle objects. Thus, the present Theorem 1 follows immediately from that result and Theorems 8 and 9. Proving Theorem 1 amounts to characterizing the brittle and semibrittle diconnected digraphs. As might be expected from the simplicity of the result, the characterization of the brittle digraphs is the easier to prove.

THEOREM 10. *Let* G *be a diconnected digraph with* $|V(G)| \geq 4$. *Then* G *is brittle if and only if* G *is dicomplete or a distar.*

Proof. It is easy to check, as we have already claimed, that dicomplete digraphs and distars are indeed brittle. Now suppose that G is brittle but not dicomplete. In order to prove that G is a distar, it will be enough to prove that it has no directed path of length 3. This is true because a diconnected digraph having at least 4 vertices and having a directed cycle of length more than 2 has a directed path of length more than 2, a diconnected digraph having no directed cycle of length more than 2 must be symmetric, and any symmetric digraph having no directed path of length more than 2 must be a distar.

The following observation will be very useful in what follows. Let v_0, v_1, v_2, v_3 be a directed path of length 3. Then we can find a split $\{V_1, V_2\}$ of G such that (v_0, v_1), $(v_2, v_3) \in \delta(V_1)$. It follows that (v_0, v_3), $(v_2, v_1) \in E(G)$.

Now let us show that the vertex-set of any directed path of length 3 is contained in a dicomplete subgraph of G. By the above remark, if the path is v_0, v_1, v_2, v_3, we have $(v_0, v_3), (v_2, v_1) \in E(G)$. There is an edge (v_3, b) for some $b \in V(G)$. If $b \neq v_1$, then we can use (v_3, b) together with (v_0, v_1) or (v_2, v_1) and an appropriate split to conclude that $(v_3, v_1) \in E(G)$. Similarly, there is an edge (c, v_0) and we can use it to obtain $(v_2, v_0) \in E(G)$. We can now use (v_1, v_2), (v_0, v_3) to conclude (v_1, v_3), $(v_0, v_2) \in E(G)$, we can use (v_3, v_1), (v_2, v_0) to conclude $(v_3, v_0) \in E(G)$, and we can use (v_3, v_0) and (v_1, v_2) to conclude (v_1, v_0), $(v_3, v_2) \in E(G)$. Therefore we have the required dicomplete subgraph.

We have seen that the existence of a directed path of length 3 implies the existence of a dicomplete subgraph H having at least 4 vertices. Choose such an H with $|V(H)|$ as large as possible. By assumption $H \neq G$ and so we can choose $(a, b) \in \delta(V(H))$. Now b is in a directed path of length 3 with every vertex of H, so (b, c), $(c, b) \in E(G)$ for every $c \in V(H)$, contradicting the choice of H. Hence, if G is not dicomplete, then G has no directed path of length 3 and so G is a distar, as required. □

THEOREM 11. *Let G be a diconnected digraph with $|V(G)| \geqq 4$. Then G is semibrittle if and only if G is a circle of transitive tournaments.*

Proof. (Throughout this proof, we abbreviate $\{W, V(G)\setminus W\}$ to $\{W, -\}$.) First, we must show that every CTT is semibrittle. Let G be a CTT with $V(G) = \{v_0, v_1, \cdots, v_{n-1}\}$ as in the definition. It is easy to see that the partitions required to be splits for G to be semibrittle are indeed splits. Now suppose that $\{V_1, V_2\}$ is an additional split. Choose $v_i, v_j \in V_1$ such that $v_k \in V_2$ for all k, $i < k < j$. We can easily check the case $|V(G)| = 4$ separately, and otherwise, either $V_1 = \{v_i, v_j\}$ or $\{\{v_i, v_{i+1}, \cdots, v_j\}, -\}$ is a split. Applying the intersection property in the latter case, we conclude that $\{\{v_i, v_j\}, -\}$ is a split. Then, using the edges (v_i, v_{i+1}), (v_j, v_{j+1}), we conclude that (v_i, v_{j+1}), $(v_j, v_{i+1}) \in E(G)$. This contradicts the assumption that G is a CTT.

The proof of the "only if" part is broken into a sequence of smaller results. We assume for convenience that $V(G) = \{1, 2, \cdots, n\}$ where $1, 2, \cdots, n$ is the semibrittle sequence and any arithmetic is modulo n (except that we use n rather than 0). Also we use the symbols $<$, $\leqq$ in a slightly nonstandard way, in order to indicate properties of betweenness. Thus, since we have in mind a circular order, $i \leqq j$ imposes no restriction on i and j, whereas $i < j$ simply means $i \neq j$. However, $i \leqq j \leqq k \leqq i$ means that $(j - i) \bmod n \leqq (k - i) \bmod n$. (The last $\leqq$ is used in the ordinary sense.)

CLAIM 1. *For each i, either $(i, i+1)$, $(i+1, i+2) \in E(G)$, or $(i+1, i)$, $(i+2, i+1) \in E(G)$.*

Proof of Claim 1. $\{\{i, i+2\}, -\}$ is not a split. Therefore, we assume that there exist (i, b), $(i+2, c) \in \delta(\{i, i+2\})$ such that not both of (i, c), $(b, i+2) \in E(G)$. (The other case is similar.) Suppose $i+1 \notin \{b, c\}$. (This implies $|V(G)| \geqq 5$.) Then $\{\{i, i+1, i+2\}, -\}$ is a split and (i, c), $(i+2, b) \in \delta(\{i, i+1, i+2\})$ but not both of (i, b), $(i+2, c) \in E(G)$, a contradiction.

Therefore, we may assume that $i+1 \in \{b, c\}$. We assume that $i+1 = b$. (The other case is similar.) We wish to prove that $(i+1, i+2) \in E(G)$. There is an edge $(d, i+2) \in E(G)$. If $d = i+1$, we are done. Otherwise, if $d \neq i$, then because $\{\{i+1, i+2\}, -\}$ is a split, we have $(i, i+2) \in E(G)$. Therefore, we may assume that $d = i$. There is an edge $(i+1, e) \in E(G)$. We may assume that $e = i$, because otherwise the split $\{\{i, i+1\}, -\}$ gives $(i+1, i+2) \in E(G)$. Now $(i+1, i)$, $(i+2, c) \in E(G)$ imply $(i+1, c) \in E(G)$. Finally, $(i, i+2)$, $(i+1, c)$ imply $(i+1, i+2) \in E(G)$. The proof of Claim 1 is complete.

CLAIM 2. *We can assume that $(i, i+1) \in E(G)$ for each i.*

Proof of Claim 2. Choose a longest directed path P of the form $i, i+1, \cdots, k$. By Claim 1, reversing the semibrittle ordering if necessary, we have $k - i \geqq 2$. If Claim

2 is not true, then by Claim 1 and the choice of P, we have that $(k+1, k)$, $(k, k-1)$, $(i+1, i)$ and $(i, i-1) \in E(G)$. Now we can conclude that $k-i \geqq 3$, because otherwise $k+1, k, i, i-1$ is a longer directed path (with respect to the reversed semibrittle ordering). Now using the split $\{\{i+1, \cdots, k-1\}, -\}$ and the edges $(i, i+1)$, $(k, k-1)$ we conclude that $(k, i+1) \in E(G)$. We can conclude from Claim 1 that $(k+2, k+1) \in E(G)$, for otherwise $(k+1, k+2)$, $(k, k+1) \in E(G)$, contradicting the maximality of P. Now using the split $\{\{k, k+1\}, -\}$ and the edges $(k-1, k)$, $(k+2, k+1)$, we obtain $(k-1, k+1) \in E(G)$. Finally, we use the split $\{\{k-1, k\}, -\}$ and the edges $(k-1, k+1)$, $(k, i+1)$ to obtain $(k, k+1) \in E(G)$. This contradicts the choice of P, and Claim 2 is proved.

CLAIM 3. *If* $(i, j) \in E(G)$, *then* $(k, l) \in E(G)$ *whenever* $i \leqq k < l \leqq j$.

Proof of Claim 3. If $i = k$, $l = j$, the result is clearly true. If $i = k$, $l < j$, then the case $l = i+1$ is clear and otherwise we can use the split $\{\{i, \cdots, l-1\}, -\}$ and the edges (i, j), $(l-1, l)$ to obtain $(i, l) \in E(G)$. Similarly, we can handle the case $k > i$, $j = l$. Finally, if $i < k < l < j$, we can use the edges (i, l), (k, j) and the split $\{\{i, i+1, \cdots, k\}, -\}$ to obtain $(k, l) \in E(G)$. This completes the proof of Claim 3.

CLAIM 4. *G has no dicomplete subgraph on* 3 *vertices.*

Proof of Claim 4. Let H be a maximal dicomplete subgraph of G, and suppose that $V(H) = \{i_1, i_2, \cdots, i_l\}$ such that $i_1 < i_2 < \cdots < i_l < i_1$ and $l \geqq 3$. Since G is not brittle, we can choose $p \in V(G) \setminus V(H)$, and we may assume $i_l < p < i_1$. Applying Claim 3 to the edge (i_l, i_{l-1}), we conclude that $(p, i_k) \in E(G)$ whenever $1 \leqq k < l$. Because $l \geqq 3$, we can use the split $\{\{p, p+1, \cdots, i_1\}, -\}$ and the edges (p, i_2) and (i_1, i_l) to obtain that $(p, i_l) \in E(G)$. A similar argument shows that $(i_k, p) \in E(G)$ for $1 \leqq k \leqq l$, and so the maximality of H is contradicted. Claim 4 is proved.

An edge $(i, j) \in E(G)$ is said to be *extreme* if there does not exist $(k, l) \in E(G)$ with $k \leqq i < j \leqq l$.

CLAIM 5. *Let* (i, j), (i', j') *be distinct extreme edges of G. The following are not possible*:

(a) $i \leqq i' < j \leqq j' < i$;
(b) $i < i' < j' < j < i$;
(c) $i < j' < i' < j < i$.

Proof of Claim 5. In case (a), suppose first that all 4 vertices are distinct. Then the edges (i, j), (i', j') and the split $\{\{i, i+1, \cdots, i'\}, -\}$ imply that $(i, j') \in E(G)$. But this contradicts the fact that (i, j) is extreme. Now suppose that $i = i'$ (so $j \neq j'$); then $(i, j') \in E(G)$, contradicting the extremeness of (i, j). Similarly, if $j = j'$, then $(i, j') \in E(G)$, contradicting the extremeness of (i', j'). Therefore (a) is impossible.

In case (b), it is clear that the extremeness of (i', j') is contradicted, so (b) is impossible.

In case (c), using (i, j) in Claim 3, we conclude that (i, i'), $(i', j) \in E(G)$. Applying Claim 3 to (i', j'), we conclude that (i', i), (j, i), $(j, j') \in E(G)$. Now using the edges (j, j') and (i, i') and the split $\{\{j, j+1, \cdots, i\}, -\}$, we conclude that $(j, i') \in E(G)$. Therefore there is a dicomplete subgraph on vertices i, j, i', contradicting Claim 4. Hence (c) is impossible, and Claim 5 is proved.

It follows from Claim 5 that pairs (i, j), (i', j') of distinct extreme edges can be of at most two kinds:

normal: $i < j \leqq i' < j' \leqq i$;
special: $j < i = j' < i' < j$ (or $j' < i' = j < i < j'$).

If all pairs of distinct extreme edges are normal, it is easy to see that G is a CTT (of type other than (n)). (Notice that there must exist at least two extreme edges.) Thus, the following result will complete the proof of Theorem 11.

CLAIM 6. *If G has a special pair of extreme edges, then G is a* CTT *of type* (n).

Proof of Claim 6. If $i+1=i'$, then $(i+1, i)\in E(G)$. If $i+1\neq i'$, then we can apply Claim 3 to the edge (i, j) to conclude that $(i+1, j)\in E(G)$. Now the split $\{\{i+1, \cdots, i'\}, -\}$ and the edges $(i+1, j)$ and $(i', i)=(i', j')$ imply that $(i+1, i)\in E(G)$. Thus, $(i+1, i)\in E(G)$ in either case. A similar argument shows that $(i, i-1)\in E(G)$. Applying Claim 3 to these two edges implies $E(G)\supseteq\{(k, l): i\leqq k<l\leqq i\}$. Therefore, G has a CTT of type (n) as a subdigraph. But adding any edge to such a CTT creates a dicomplete subgraph on 3 vertices. Therefore, by Claim 4, G is a CTT of type (n) and Claim 6 is proved. The proof of Theorem 11 is complete. □

4. Algorithms. In this section we describe polynomial-time algorithms for testing a diconnected digraph for primeness and for the apparently harder problem of constructing its standard decomposition. Such algorithms have been discovered for the case of the substitution decomposition for undirected graphs [2], [4], [8], and for digraphs [10]. Transitive acyclic digraphs can also be treated as a special case of undirected graphs (see [2]). To my knowledge these are the only cases which have been previously solved.

We summarize the properties of the algorithms of this section in the following results. In their statements, and throughout the section, we let n denote $|V(G)|$ and m denote $|E(G)|$.

THEOREM 12. *Suppose that G is diconnected. There is an algorithm requiring $O(n^4)$ time and $O(m)$ space to do any of the following*:

(a) *Determine whether or not G is prime*;

(b) *Compute a prime decomposition of G*;

(c) *Compute the standard decomposition of G.*

Moreover, if G is either pointed or symmetric, there is an algorithm requiring $O(n^3)$ time to do any of (a), (b), (c).

COROLLARY 2. *If G is any digraph, there is an algorithm to compute the factorization of Theorem* 5, *and hence test G for irreducibility, in $O(n^3)$ time and $O(n+m)$ space.*

Thus the decomposition algorithms corresponding to the classes of digraphs in Theorems 3 through 6 are all $O(n^3)$; it is only for the most general decomposition of Theorem 2 that an $O(n^4)$ algorithm is required. In the cases in which algorithms have already been known (substitution decomposition) these results improve the bound [10] for digraphs and match those for special classes [2], [8].

We are ready to begin describing the algorithms. The basic problem to be solved is to find, if possible, a split of a given diconnected digraph G. It is convenient to attack first the following more restricted problems.

Problem 1. Given edges (x_1, y_1), (x_2, y_2) of G, find, if there is one, a split $\{V_1, V_2\}$ of G such that $x_1, y_2\in V_1$ and $x_2, y_1\in V_2$.

Problem 2. Given edges (x_1, y_1), (x_2, y_2) of G and a set $S\subseteq V(G)$ satisfying $x_1, y_2\in S$, $x_2, y_1\notin S$ and $|S|\geqq 2$, find, if there is one, a split $\{V_1, V_2\}$ of G such that $x_2, y_1\notin V_1\supseteq S$.

It is easy to see that an efficient algorithm for Problem 1 will yield an efficient algorithm for the fundamental problem of finding a split of G. We will describe an efficient algorithm for Problem 2, and show that this algorithm can be used to solve Problem 1. The algorithm for Problem 2 is based on Proposition 4 below, which provides a more economical way to recognize splits. We will make use of some terminology. If $(x, y)\in E(G)$ and $p, q\in V(G)$, we say that $P(x, y, p, q)$ holds if the following condition fails:

$$(p, q)\in E(G) \quad \text{if and only if} \quad (p, y), (x, q)\in E(G).$$

PROPOSITION 4. *Let G be a diconnected digraph, let $S \subseteq V(G)$ such that $|S| \geqq 2 \leqq |V(G)\backslash S|$ and let $(x_1, y_1) \in \delta(S)$, $(x_2, y_2) \in \delta(V(G)\backslash S)$. Then $\{S, V(G)\backslash S\}$ is a split of G if and only if there does not exist $p \in S$, $q \in V(G)\backslash S$ such that $P(x_1, y_1, p, q)$ or $P(x_2, y_2, q, p)$.*

Proof. The "only if" part is obvious. For the "if" part, suppose that $\{S, V(G)\backslash S\}$ is not a split. By symmetry, we may assume that there exist (a, b), $(c, d) \in \delta(S)$ such that $(a, d) \notin E(G)$. If neither $P(x_1, y_1, a, b)$ nor $P(x_1, y_1, c, d)$ holds, then (a, y_1), (x_1, b), (c, y_1), $(x_1, d) \in E(G)$. If $P(x_1, y_1, a, d)$ holds, then at least one of (x_1, d), (a, y_1) is not an edge of G, a contradiction. □

If we are attempting to solve Problem 2, and we discover $p \in S$, $q \in V(G)\backslash S$ such that $P(x_1, y_1, p, q)$ or $P(x_2, y_2, q, p)$, then any solution $\{V_1, V_2\}$ must satisfy $q \in V_1$, and so it is equivalent to solve Problem 2 with S replaced by $S \cup \{q\}$. On the other hand, if no such p, q exist then $\{S, V(G)\backslash S\}$ is a split (provided only that $|V(G)\backslash S| \geqq 2$). These remarks lead naturally to an algorithm to solve Problem 2.

ALGORITHM 1. (Input is as described in Problem 2.)

```
begin
    T := S;
    while T ≠ ∅ do
        Select p ∈ T; T := T\{p};
        for q ∈ V(G) do
            if q ∉ S and [P(x1, y1, p, q) or P(x2, y2, q, p)] then
                S := S ∪ {q}; T := T ∪ {q};
            endif
        endfor
    endwhile
end
```

THEOREM 13. *Suppose that Algorithm 1 is applied to a diconnected digraph G. If the algorithm terminates with one of $x_2, y_1 \in S$ or with $|S| = n - 1$, then there is no split of the kind required in Problem 2, and otherwise $\{S, V(G)\backslash S\}$ is such a split. Moreover, Algorithm 1 can be implemented in $O(n^2)$ time and $O(m)$ space.*

Proof. By Proposition 4 and the remark following it, any algorithm which adds to S successively elements q of $V(G)\backslash S$ such that, for some $p \in S$, $P(x_1, y_1, p, q)$ or $P(x_2, y_2, q, p)$ will have the termination property stated in the theorem. For the sake of efficiency, Algorithm 1 performs these operations in a special order. Namely, it checks for a given $p \in S$ and every q currently in $V(G)\backslash S$ whether q should be added to S, and then never uses p in this way again. The set T is used to make sure that every legal choice for p is eventually used. This proves the validity of Algorithm 1, as claimed in the statement of the theorem.

Now we consider the efficiency of the algorithm. To meet the desired $O(m)$ space bound, we represent the digraph by keeping, for each $v \in V(G)$, two lists, the "out-list" consisting of all w such that $(v, w) \in E(G)$ and the "in-list" consisting of all u such that $(u, v) \in E(G)$. (Certainly this representation of G can be constructed within $O(n^2)$ time from any of the usual input representations of G.) We keep the set T as a list and the set S as a characteristic vector. Thus we can perform in constant time each of the individual operations on S and T required by the algorithm. If we have characteristic vectors for the out-list of each of x_1, x_2 and p and for the in-list of each of y_1, y_2 and p, then it is easy to see that each application of the **if** statement requires only constant time. Therefore, each execution of the **for** statement can be done in $O(n)$ time. We create the characteristic vectors for the in-list and out-list of p when

p is selected from T and discard these vectors when we return to select a new p. Therefore, using these vectors does not affect the $O(m)$ space bound. Moreover, the total work to construct these vectors is $O(n)$ for each p. Since there are $O(n)$ choices for p, the desired $O(n^2)$ time bound for Algorithm 1 is established. □

It is quite easy to use Algorithm 1 to solve Problem 1. If $x_1 \neq y_2$, we can begin Algorithm 1 with $S = \{x_1, y_2\}$. If $x_1 = y_2$, but $x_2 \neq y_1$, we can interchange the roles of (x_1, y_1), (x_2, y_2) and proceed as above. The remaining case, in which $x_1 = y_2$ and $x_2 = y_1$, is a little more difficult. We choose a vertex z different from x_1 and y_1 and run Algorithm 1 twice. The first time S is initialized to be $\{x_1, z\}$, and the second time (x_1, y_1) and (x_2, y_2) are interchanged and S is initialized to be $\{y_1, z\}$. The first application of Algorithm 1 tests for a solution $\{V_1, V_2\}$ of Problem 1 such that $z \in V_1$, and the second tests for $\{V_1, V_2\}$ with $z \in V_2$. Therefore, in all cases we have an $O(n^2)$ algorithm for Problem 1.

To test a diconnected digraph for primeness, we can clearly solve Problem 1 m^2 times, giving an $O(n^2m^2)$ algorithm. However, this bound can be considerably improved. We choose a vertex $r \in V(G)$ and construct a spanning out-tree T_1 of G rooted at r, and a spanning in-tree T_2 of G rooted at r. Since G is diconnected, both T_1 and T_2 must exist, and they can be constructed in $O(m)$ time. Now, given a split $\{V_1, V_2\}$ of G, we may assume that $r \in V_1$, and hence $\delta(V_1)$ contains at least one edge from T_1 and $\delta(V_2)$ contains at least one edge from T_2. Therefore, we can test G for primeness by solving Problem 1 repeatedly, where (x_1, y_1) runs through $E(T_1)$ and (x_2, y_2) runs through $E(T_2)$. This requires $O(n^2)$ instances of Problem 1, and thus we have an algorithm for finding a split or proving primeness which requires $O(n^4)$ time and $O(m)$ space.

In the two important special cases of the digraph decomposition, the above approach simplifies and becomes more efficient. Suppose that G has the property that the spanning trees T_1, T_2 can be chosen so as to satisfy $E(T_2) = \{(x, y): (y, x) \in E(T_1)\}$. Then for any split $\{V_1, V_2\}$ with $r \in V_1$, $\delta(V_1)$ contains at least one element (x, y) of $E(T_1)$, and then $(y, x) \in \delta(V_2) \cap E(T_2)$. Therefore, it will be sufficient in this case to solve Problem 1 $n-1$ times, with (x_1, y_1) running through $E(T_1)$ and (x_2, y_2) always equal to (y_1, x_1). It follows that, when T_1, T_2 can be chosen in this special way, the time bound for prime-testing becomes $O(n^3)$. Of course, if G is symmetric, we can find such T_1, T_2; namely, they arise from any spanning tree of the undirected graph corresponding to G. Similarly, if the digraph G is pointed, then the edges incident with the point provide such T_1, T_2. It follows that we have an $O(n^3)$ algorithm to test a digraph for irreducibility.

We have proved the parts of Theorem 12 that concern prime recognition. Next we explain how an algorithm for constructing a prime decomposition of G leads to an algorithm for constructing the standard decomposition of G. Clearly, it will be enough to be able to recognize whether a decomposition D consisting of primes, brittles and semibrittles is minimal with this property; if it is not, we will construct another such decomposition D' of which D is a strict refinement. We need to recognize whether two members G_1, G_2 of D, sharing a marker v, comprise a simple decomposition of a brittle or semibrittle digraph. This can be done with the aid of the following result. We omit its (straightforward) proof. We also leave to the reader the task of verifying that the additional work required to use Proposition 5 to construct the standard decomposition from a prime decomposition does not violate the time and space bounds of Theorem 12.

PROPOSITION 5. *Let G_1, G_2 be diconnected digraphs such that $V(G_1) \cap V(G_2) = \{v\}$ and $|V(G_1)| \geqq 3 \leqq |V(G_2)|$. Then $G = G_1 * G_2$ is brittle or semibrittle if and only if*

one of the following is true:

(a) G_1, G_2 *are both dicomplete*;

(b) G_1, G_2 *are both distars and* v *is the center of exactly one of them*;

(c) G_1, G_2 *are both* CTT's *and* v *is a hinge of both of them.*

Finally, we explain how the algorithms for finding a split lead to equally efficient algorithms for constructing prime decompositions. (Similar ideas are used in a different application [7] to show that algorithms for prime-testing and constructing prime decompositions have the same complexity.) Let r be a vertex of the diconnected digraph G, and let T be a spanning out-tree rooted at r. Every split $\{V_1, V_2\}$ of G with $r \in V_1$ satisfies $\delta(V_1) \cap E(T) \neq \emptyset$. For each edge $(x, y) \in E(T)$ we can test in time $O(n^3)$ for the existence of a split $\{V_1, V_2\}$ of G with $x, r \in V_1$ and $y \in V_2$. Suppose that we find such a split and construct the resulting simple decomposition $\{G_1, G_2\}$ of G with marker v. In searching for splits in G_1 and G_2, it is possible to take advantage of work already expended on G. In order to do this, we use spanning trees in G_1 and G_2 which are constructed directly from T.

Let (a, b) be an edge of T such that $a \in V_1$, $b \in V_2$ and (a, b) is the first such edge of a directed path in T from r to a vertex in V_2. We define spanning out-trees T_1 of G_1 rooted at r and T_2 of G_2 rooted at v, as follows: $E(T_1) = \{(a, v)\} \cup \{(x, y): (x, y) \in E(T)$ and $x, y \in V_1\} \cup \{(v, y): y \in V_1$ and for some $x \in V_2$, $(x, y) \in E(T)\}$; $E(T_2) = \{(x, y): (x, y) \in E(T)$ and $x, y \in V_2\} \cup \{(v, y): y \in V_2$ and for some $x \in V_1$, $(x, y) \in E(T)\}$. It is not difficult to see that T_1, T_2 are indeed spanning directed out-trees of G_1, G_2 and that there is a one-to-one correspondence between the elements of the sets $E(T_1) \cup E(T_2)$ and $E(T) \cup \{(a, b)\}$.

The algorithm for constructing a prime decomposition maintains a decomposition $D = \{G_i: i \in I\}$ of G and a spanning out-tree T_i of G_i for each $i \in I$. When a split of some G_i is found, a simple decomposition is formed, new spanning trees are defined as above, and D is refined. If an edge (p, q) of some T_i is found not to yield a split of G_i (that is, there is no split $\{V_3, V_4\}$ of G_i having $p \in V_3$, $q \in V_4$), then it follows from F3 and this construction that (p, q) will be an edge of some T_j in every subsequent decomposition. Moreover, (p, q) can never yield a split of G_j so it need never be tested again. Any decomposition D of G consists of at most $n-2$ digraphs, and any collection consisting of a spanning tree of each of the members of D has a total of at most $2n-4$ edges. Therefore, the total number of times that edges are tested for yielding a split is $O(n)$. (Of course, many of these tests will be done on digraphs smaller than G.) Therefore, the running time for the entire algorithm is $O(n^4)$. In the two special cases mentioned in Theorem 12, each edge can be tested in $O(n^2)$ leading to a running time of $O(n^3)$. The proof of Theorem 12 is complete.

Acknowledgment. My interest in this subject is due largely to my collaboration with Jack Edmonds on decomposition theory and to conversations with Michel Habib on graph decomposition. I am grateful to both of them. I am also indebted to J. Tan for a number of corrections and improvements.

REFERENCES

1. J. A. BONDY AND U. S. R. MURTY, *Graph Theory with Applications*, Macmillan, London, 1976.
2. H. BUER AND R. MOHRING, *A fast algorithm for the decomposition of graphs and posets*, preprint, 1980.
3. V. CHVÁTAL, *On certain polytopes associated with graphs*, J. Combin. Theory B, 18 (1975), pp. 138–154.
4. D. D. COWAN, L. O. JAMES AND R. G. STANTON, *Graph decomposition for undirected graphs*, in Proc. 3rd Southeastern Conference, Hoffman and Levow, eds. Utilitas Mathematica, Winnipeg, Manitoba, Canada, 1972.

5. W. H. CUNNINGHAM, *A combinatorial decomposition theory*, Thesis, University of Waterloo, Waterloo, Ontario, Canada, 1973.
6. W. H. CUNNINGHAM AND J. EDMONDS, *A combinatorial decomposition theory*, Canad. J. Math., 32 (1980), pp. 734–765.
7. ——— *Decomposition of linear systems*, in preparation.
8. M. HABIB AND M. C. MAURER, *On the X-join decomposition for undirected graphs*, Discrete Appl. Math. 1 (1979), pp. 201–207.
9. M. C. MAURER, *Unité de la décomposition d'un graphe en joint suivant un graphe joint-irréductible, d'une famille de ses sous-graphes*, C. R. Acad. Sci. Paris, 283 (1976), pp. 289–292.
10. ——— *Joints et décompositions premières dans les graphes*, Thèse 3 ème cycle, Universitè Paris VI, 1977.
11. G. SABIDUSSI, *Graph derivatives*, Math. Zeitschr., 76 (1961), pp. 385–401.
12. J. B. SIDNEY, *Decomposition algorithms for single-machine sequencing with precedence relations and deferral costs*, Oper. Res. 23 (1975), pp. 283–298.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 2, June 1982

0196-5212/82/0302-0009 $01.00/0

ON THE PROBLEM OF PARTITIONING PLANAR GRAPHS*

HRISTO NICOLOV DJIDJEV†

Abstract. The results in this paper are closely related to the effective use of the divide-and-conquer strategy for solving problems on planar graphs. It is shown that every planar graph can be partitioned into two or more components of roughly equal size by deleting only $O(\sqrt{n})$ vertices, and such a partitioning can be found in $O(n)$ time. Some of the theorems proved in the paper are improvements on the previously known theorems while others are of more general form. An upper bound for the minimum size of the partitioning set is found.

1. Introduction. Many kinds of combinatorial problems can be solved efficiently using the method "divide-and-conquer" [1]. In this method the original problem is divided into two or more smaller problems, each of the subproblems is solved by applying the same method recursively, and the solutions to the subproblems are finally combined to give the solution to the original problem. In [3] the next three conditions are shown to be necessary for the success and efficiency of divide-and-conquer: (i) the subproblems must be of the same type as the original and independent of each other (in a suitable sense); (ii) the cost of combining the subproblem solutions into a solution to the original problem must be small; and (iii) the subproblems must be substantially smaller than the original problem.

For problems defined on graphs we offer more general conditions under which the divide-and-conquer approach is useful. Let S be a class of graphs closed under the subgraph relation (i.e., if $G_1 \in S$ and G_2 is a subgraph of G_1, then $G_2 \in S$). In [3] an *$f(n)$-separator theorem* for S is defined as a theorem of the following form:

THEOREM A. *There exist constants $\alpha < 1$, $\beta > 0$ such that if G is any n-vertex graph in S, then the vertices of G can be partitioned into three sets A, B, C such that no edge joins a vertex in A with a vertex in B, $|A| \leqq \alpha n$, $|B| \leqq \alpha n$, $|C| \leqq \beta f(n)$.*

If G is the graph in S on which the problem is defined, then the subgraphs induced by the sets of vertices A and B define subproblems, which are relatively independent of each other. The cost of combining the solutions to the subproblems into a solution to the original problem depends on the number of vertices in C (and thus on $f(n)$). So if there exists a fast algorithm for finding the appropriate vertex partition A, B, C and $f(n) = o(n)$, then Theorem A makes possible the use of divide-and-conquer for solving different problems defined on graphs in S.

Lipton and Tarjan [3] proved that a $\sqrt{n}$-separator theorem holds for the class of all planar graphs. In this paper some improvements are made on their results. The most important result in [3] is a $\sqrt{n}$-separator theorem with $\alpha = \frac{2}{3}$ and $\beta = 2\sqrt{2}$. In § 2 of this paper a similar theorem is proved with $\alpha = \frac{2}{3}$ and $\beta = \sqrt{6}$. In § 3 it is shown that no $\sqrt{n}$-separator theorem holds for the class of all planar graphs for $\alpha = \frac{2}{3}$ and $\beta < \sqrt{4\pi\sqrt{3}}/3$. In § 4 we prove theorems of different form from that offered above, and a planar separator theorem for $\alpha = \frac{1}{2}$ and $\beta = (3+\sqrt{21}/2)+3\sqrt{2}/(\sqrt{3}-1)$.

2. We shall make use of the next statement, proved in [3]:

LEMMA 1. *Let G be any n-vertex connected planar graph. Suppose that the vertices of G are partitioned into levels according to their distance from some vertex v, and that $L(l)$ denotes the number of vertices on level l. Given any two levels l_1 and l_2 such that the number of vertices on levels 0 through $l_1 - 1$ does not exceed $2n/3$ and the number*

* Received by the editors November 10, 1980.
† Faculty of Mathematics, Sofia University, Sofia, Bulgaria.

of vertices on levels l_2+1 and above does not exceed $2n/3$, it is possible to find a partition A, B, C of the vertices of G such that no edge joins a vertex in A with a vertex in B, $|A|\leqq 2n/3$, $|B|\leqq 2n/3$, $|C|\leqq L(l_1)+L(l_2)+\max\{0, 2(l_2-l_1-1)\}$.

THEOREM 1. *Let G be any n-vertex planar graph. The vertices of G can be partitioned into three sets A, B, C such that no edge joins a vertex in A with a vertex in B, $|A|\leqq 2n/3$, $|B|\leqq 2n/3$, $|C|\leqq\sqrt{6n}$.*

Proof. Assume G is connected. Partition the vertices into levels according to their distance from some vertex v. Let $L(l)$ be the number of vertices on level l. If r is the maximum distance of any vertex from v, define additional levels -1 and $r+1$ containing no vertices.

For each $\alpha\in(0, 1)$ let l_α denote a level such that

$$\sum_{l=0}^{l_\alpha-1} L(l)<\alpha n, \qquad \sum_{l=0}^{l_\alpha} L(l)\geqq\alpha n.$$

Case 1. There exists a level l such that $l_{1/3}\leqq l\leqq l_{2/3}$ and $L(l)\leqq\sqrt{6n}$. Let A be the set of vertices on levels 0 through $l-1$; let B be the set of vertices on levels $l+1$ through r, and let C be the set of vertices on level l. Then the theorem is true.

Case 2. For each $l\in[l_{1/3}, l_{2/3}]$, $L(l)>\sqrt{6n}$. Let $\alpha=(\sum_{l=l_{1/2}}^{l_{2/3}} L(l))/n$. Since $\sum_{l=l_{1/3}}^{l_{2/3}} L(l)=\sum_{l=0}^{l_{2/3}} L(l)-\sum_{l=0}^{l_{1/3}-1} L(l)>\frac{2}{3}n-\frac{1}{3}n=\frac{1}{3}n$, then $\alpha>\frac{1}{3}$. Furthermore

$$\alpha n=\sum_{l=l_{1/3}}^{l_{2/3}} L(l)>\sum_{l=l_{1/3}}^{l_{2/3}}\sqrt{6n}=(l_{2/3}-l_{1/3}+1)\sqrt{6n}.$$

Thus

$$l_{2/3}-l_{1/3}+1<\frac{\alpha}{\sqrt{6}}\sqrt{n}. \tag{1}$$

Let j be a nonnegative integer such that

$$\sum_{l=l_{1/3}-j+1}^{l_{2/3}+j-1} L(l)<\tfrac{2}{3}n, \qquad \sum_{l=l_{1/3}-j}^{l_{2/3}+j} L(l)\geqq\tfrac{2}{3}n.$$

Subcase 2.1. There exists i such that $0\leqq i\leqq j$ and $L(l_{1/3}-i)+L(l_{2/3}+i)\leqq\sqrt{6n}$. Then let C be the set of the vertices on levels $l_{1/3}-i$ and $l_{2/3}+i$, let A be the set of the vertices on levels $l_{1/3}-i+1$ through $l_{2/3}+i-1$, and let B be the set of the remaining vertices. Then the theorem is true.

Subcase 2.2. For each i, $1\leqq i\leqq j$ we have

$$L(l_{1/3}-i)+L(l_{2/3}+i)>\sqrt{6n}.$$

Let $\beta=(\sum_{l=l_{1/3}-j}^{l_{2/3}+j} L(l))/n$. Then $\beta\geqq\frac{2}{3}$. Furthermore

$$\begin{aligned}\beta n=\sum_{l=l_{1/3}-j}^{l_{2/3}+j} L(l)&=\sum_{l=l_{1/3}-j}^{l_{1/3}-1} L(l)+\sum_{l_{1/3}}^{l_{2/3}} L(l)+\sum_{l_{2/3}+1}^{l_{2/3}+j} L(l)\\&=\alpha n+\sum_{i=1}^{j}[L(l_{1/3}-i)+L(l_{2/3}+i)]>\alpha n+j\sqrt{6n}.\end{aligned}$$

Thus $(\beta-\alpha)n>j\sqrt{6n}$ and

$$j<\frac{\beta-\alpha}{\sqrt{6}}\sqrt{n}. \tag{2}$$

Let $m=\sum_{l=0}^{l_{1/3}-j-1} L(l)$. Then $\sum_{l=l_{2/3}+j+1}^{2} L(l)=n-m-\beta n$.

Using the same idea as in the proof of [3, Theorem 4], it is easy to show that there exist levels l' and l'' such that

$$l' \leqq l_{1/3} - j - 1 < l_{2/3} + j + 1 \leqq l'',$$

$$L(l') + 2(l_{1/3} - j - 1 - l') \leqq 2\sqrt{m}, \tag{3}$$

$$L(l'') + 2(l'' - (l_{2/3} + j + 1)) \leqq 2\sqrt{n - m - \beta n}. \tag{4}$$

Add the inequalities (3) and (4) to give

$$L(l') + L(l'') + 2(l'' - l' - 1 - (l_{2/3} - l_{1/3} + 2j + 1)) \leqq 2\sqrt{m} + 2\sqrt{n - m - \beta n}. \tag{5}$$

Then multiply the inequality (1) by 2, the inequality (2) by 4 and add them together. The result is

$$2(l_{2/3} - l_{1/3} + 2j + 1) < \frac{2\alpha}{\sqrt{6}}\sqrt{n} + \frac{4(\beta - \alpha)}{\sqrt{6}}\sqrt{n} = \frac{4\beta - 2\alpha}{\sqrt{6}}\sqrt{n}. \tag{6}$$

Finally add (5) and (6):

$$L(l') + L(l'') + 2(l'' - l' - 1) < 2(\sqrt{m} + \sqrt{n - m - \beta n}) + \frac{4\beta - 2\alpha}{\sqrt{6}}\sqrt{n}. \tag{7}$$

However, $\sqrt{m} + \sqrt{n - m - \beta n} \leqq \sqrt{(n - \beta n)/2} + \sqrt{(n - \beta n)/2} = \sqrt{2}\sqrt{1 - \beta}\sqrt{n}$, whence

$$L(l') + L(l'') + 2(l'' - l' - 1) < (2\sqrt{2}\sqrt{1 - \beta} + (4\beta - 2\alpha)/\sqrt{6})\sqrt{n}.$$

Since $\alpha \geqq \frac{1}{3}$ then $(4\beta - 2\alpha)/\sqrt{6} \leqq (4\beta - \frac{2}{3})/\sqrt{6}$. Let $f(x) = 2\sqrt{2}\sqrt{1 - x} + (4x - \frac{2}{3})/\sqrt{6}$, $x \in [\frac{2}{3}, 1]$. Then

$$f'(x) = -\frac{\sqrt{2}}{\sqrt{1 - x}} + \frac{4}{\sqrt{6}} = \frac{4\sqrt{1 - x} - 2\sqrt{3}}{\sqrt{6}\sqrt{1 - x}} = \frac{4}{\sqrt{6}\sqrt{1 - x}}\left(\sqrt{1 - x} - \sqrt{\frac{3}{4}}\right).$$

Then f is a decreasing function and $f(x) \leqq f(\frac{2}{3})$, $x \geqq \frac{2}{3}$. Since $\beta \geqq \frac{2}{3}$ then $f(\beta) \leqq f(\frac{2}{3})$. Thus:

$$2\sqrt{2}\sqrt{1 - \beta} + \frac{4\beta - \frac{2}{3}}{\sqrt{6}} \leqq 2 \cdot \frac{\sqrt{2}}{\sqrt{3}} + \frac{4 \cdot \frac{2}{3} - \frac{2}{3}}{\sqrt{6}} = \frac{4}{\sqrt{6}} + \frac{2}{\sqrt{6}} = \sqrt{6}.$$

Then $L(l') + L(l'') + 2(l'' - l' - 1) < \sqrt{6n}$ and by Lemma 1 the theorem is true in this subcase. This completes the proof for connected graphs.

The proof in the case when G is not connected is the same as in Theorem 4 in [3]. □

3. Now we shall find a lower bound for the smallest constant which can replace $\sqrt{6}$ in Theorem 1.

We shall use without a proof the next geometrical statement.

LEMMA 2. *From all the curves upon a given sphere which divide it into two parts, the ratio of the areas of which parts is equal to a given positive constant, the circumference has a minimum length.*

Further, the area of surface A and the length of a line L will be denoted by $S(A)$ and $l(L)$ respectively.

LEMMA 3. *Let E_1 be a sphere with radius 1 and let L be a curve upon E_1, dividing it into two parts A and B, such that $S(A) \leqq \beta S(E_1)$, $S(B) \leqq \beta S(E_1)$, where $\beta \in [\frac{1}{2}, 1]$. Then $l(L) \geqq 4\pi\sqrt{\beta - \beta^2}$.*

Proof. Suppose without loss of generality that $S(A) \leqq S(B)$ and let $\alpha = S(A)/S(B)$.

Let k_1 be a circumference, which divides E_1 into two parts A' and B' such that $S(A')/S(B') = \alpha$. Then by Lemma 2

$$l(k_1) \leqq l(L). \tag{8}$$

Obviously $S(A') = S(A)$ and $S(B') = S(B)$, whence $S(A') \leqq S(B') \leqq \beta S(E_1)$. Let k_2 be a circumference, which divides the sphere into two parts A'' and B'' such that $S(A'') = (1-\beta)S(E_1)$, $S(B'') = \beta S(E_1)$. Then apparently

$$l(k_2) \leqq l(k_1). \tag{9}$$

Let us calculate $l(k_2)$.

Present A'' as a result of the rotation of the curve $y = \sqrt{1-x^2}$, where $x \in [z, 1]$ (z unknown), round the axis Ox. Then

$$y' = \frac{-x}{\sqrt{1-x^2}}, \qquad ds = \frac{dx}{\sqrt{1-x^2}},$$

and hence

$$4\pi(1-\beta) = S(A'') = 2\pi \int_z^1 y\, ds = 2\pi \int_z^1 dx = 2\pi(1-z),$$

whence $z = 2\beta - 1$. Then

$$l(k_2) = 2\pi\sqrt{1-(2\beta-1)^2} = 4\pi\sqrt{\beta-\beta^2}. \tag{10}$$

By (8), (9) and (10) it follows that $l(L) \geqq 4\pi\sqrt{\beta-\beta^2}$. □

With each polygon P and a number $\alpha > 0$ will be associated a graph $V_\alpha(P)$ in the described manner:

Bring in the plane of the polygon three systems of parallel straight lines, which divide it into equilateral triangles with sides α (Fig. 1).

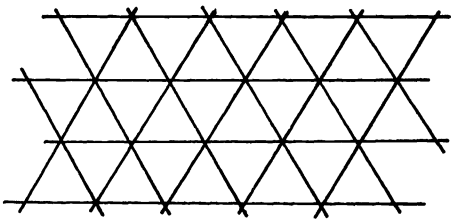


FIG. 1

To the resulting nest add the contour of the polygon. Finally, for vertices of $V_\alpha(P)$ define the nodes of the net, and for edges, the segments of the net connecting the nodes.

LEMMA 4. *The number of the vertices of $V_\alpha(P)$ which are adjacent to vertices of the contour of P is $O(\alpha^{-1})$.*

Proof. Let a be one of the sides of the polygon. Then the number of the points of intersection of a with each of the systems is at most $l(a)/(\alpha\sqrt{3}/2)+1$, and at most $3[l(a)/(\alpha\sqrt{3}/2)+1]$ with the whole nest. Each of these points is adjacent to no more than 8 vertices of $V_\alpha(P)$. Therefore the number of vertices of $V_\alpha(P)$ which are adjacent to vertices of a, and thus to vertices of the whole contour of P is $O(\alpha^{-1})$. □

Let n_α be the number of the vertices of $V_\alpha(P)$.

LEMMA 5.

$$\lim_{\alpha\to 0} n_\alpha \cdot \alpha^2 = (2/\sqrt{3})S(P).$$

Proof. Let $\tilde{n}_\alpha$ be the number of the vertices of $V_\alpha(P)$, which are adjacent to vertices of the contour of P. Then by Lemma 4 $\tilde{n}_\alpha = O(\alpha^{-1})$.

Circumscribe a rhomb with side α around each of the vertices of $V_\alpha(P)$ as shown in Fig. 2. These rhombs cover the polygon P.

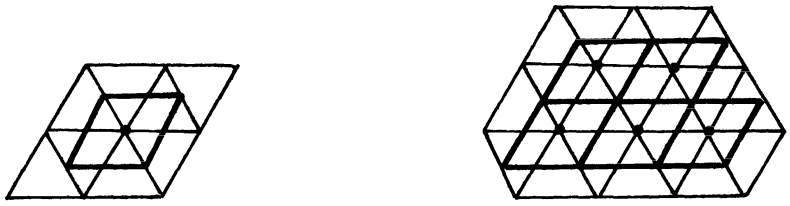


FIG. 2

The area of each rhomb is equal to $(\sqrt{3}/2)\alpha^2$. Hence

$$(n_\alpha - \tilde{n}_\alpha)\sqrt{\tfrac{3}{2}}\alpha^2 \leqq S(P) \leqq n_\alpha \cdot \sqrt{\tfrac{3}{2}}\alpha^2,$$

$$n_\alpha \cdot \alpha^2 - \tilde{n}_\alpha \cdot \alpha^2 \leqq \frac{2}{\sqrt{3}} S(P) \leqq n_\alpha \cdot \alpha^2, \tag{11}$$

$$0 \leqq n_\alpha \cdot \alpha^2 - \frac{2}{\sqrt{3}} S(P) \leqq \tilde{n}_\alpha \cdot \alpha^2.$$

Since $\tilde{n}_\alpha = O(\alpha^{-1})$ then

$$\lim_{\alpha\to 0} \tilde{n}_\alpha \cdot \alpha^2 = 0. \tag{12}$$

By (11) and (12) it follows that

$$\lim_{\alpha\to 0} n_\alpha \cdot \alpha^2 = \frac{2}{\sqrt{3}} S(P). \tag{13}$$

□

THEOREM 2. *The smallest constant which can replace* $\sqrt{6}$ *in Theorem* 1 *must be no smaller than* $(\sqrt{4\pi\sqrt{3}})/3 \approx 1.555$.

Proof. Let E_1 be a sphere with radius 1 and center the point O, and let M_1, $M_2, \cdots, M_n, \cdots$ be a sequence of convex polyhedrons such that

$$r(M_n, E_1) \xrightarrow[n\to\infty]{} 0. \tag{14}$$

where $r(\cdot, \cdot)$ denotes the Hausdorff distance in $\mathbb{R}^3$. To each side of these polyhedrons, treated as a polygon, and each $\alpha > 0$ corresponds a graph (as described above). Thus to each polyhedron M_n and each $\alpha > 0$ corresponds a planar graph $V_\alpha(M_n)$ (or briefly $V_{\alpha,n}$).

Let $A_{\alpha,n}$, $B_{\alpha,n}$, $C_{\alpha,n}$ be a partition of the vertices of $V_{\alpha,n}$ such that no edge joins a vertex in $A_{\alpha,n}$ with a vertex in $B_{\alpha,n}$, and each of $A_{\alpha,n}$ and $B_{\alpha,n}$ contains no more than $2n_\alpha/3$ vertices, where n_α is the number of vertices in $V_{\alpha,n}$.

Let $\tilde{C}_{\alpha,n}$ be the set of the edges in $V_{\alpha,n}$, both endpoints of which belong to $C_{\alpha,n}$. $\tilde{C}_{\alpha,n}$ divides M_n into two regions $\tilde{A}_{\alpha,n}$ and $\tilde{B}_{\alpha,n}$ with contour $\tilde{C}_{\alpha,n}$, containing the vertices of $A_{\alpha,n}$ and $B_{\alpha,n}$ respectively. By Lemma 5,

$$S(\tilde{A}_{\alpha,n}) \leqq \tfrac{2}{3}S(M_n) + \varepsilon_1(n,\alpha), \tag{15}$$

$$S(\tilde{B}_{\alpha,n}) \leqq \tfrac{2}{3}S(M_n) + \varepsilon_1(n,\alpha), \tag{16}$$

where $\varepsilon_1(n,\alpha) \xrightarrow[\alpha\to 0]{} 0$.

Let $\varphi \cdot \mathbb{R}^3 \backslash O \to E_1$ such that for $x \in \mathbb{R}^3 \backslash O$, $\varphi(x)$ belongs to the ray Ox. By (14), (15) and (16)

$$S(\varphi(\tilde{A}_{\alpha,n})) \leqq \tfrac{2}{3}S(E_1) + \varepsilon_2(n,\alpha), \tag{17}$$

$$S(\varphi(\tilde{B}_{\alpha,n})) \leqq \tfrac{2}{3}S(E_1) + \varepsilon_2(n,\alpha), \tag{18}$$

where $\varepsilon_2(n,\alpha) \xrightarrow[n\to\infty,\alpha\to 0]{} 0$. By (17), (18) and Lemma 3 it follows that $\varphi(\tilde{C}_{\alpha,n})$ contains a curve $L'_{\alpha,n}$ such that

$$l(L'_{\alpha,n}) \geqq \frac{4\sqrt{2}\pi}{3} + \bar{\varepsilon}(n,\alpha),$$

where $\bar{\varepsilon}(n,\alpha) \xrightarrow[n\to\infty,\alpha\to 0]{} 0$, since $4\pi\sqrt{\beta-\beta^2}|_{\beta=2/3} = 4\sqrt{2}\pi/3$. By (14)

$$l(L_{\alpha,n}) \geqq \frac{4\sqrt{2}\pi}{3} + \varepsilon(n,\alpha),$$

where $\varepsilon(n,\alpha) \xrightarrow[n\to\infty,\alpha\to 0]{} 0$, $L_{\alpha n} \subset \tilde{C}_{\alpha,n}$, $\varphi(L_{\alpha,n}) = L'_{\alpha,n}$. Since $S(E_1) = 4\pi$ then

$$l(L_{\alpha,n}) \geqq \frac{\sqrt{S(E_1)}\,2\sqrt{2}\,\sqrt{\pi}}{3} + \varepsilon(n,\alpha) \geqq \frac{\sqrt{2S(M_n)}\,2\sqrt{\pi}}{3} + \varepsilon(n,\alpha)$$

$$= \sqrt{\frac{2S(Mn)}{\sqrt{3}}}\sqrt{\frac{4\pi\sqrt{3}}{9}} + \varepsilon(n,\alpha) = \sqrt{n_\alpha \cdot \alpha^2} \cdot \sqrt{\frac{4\pi\sqrt{3}}{9}} + \bar{\bar{\varepsilon}}(n,\alpha),$$

where $\bar{\bar{\varepsilon}}(n,\alpha) \xrightarrow[n\to\infty,\alpha\to 0]{} 0$. Therefore

$$l(L_{\alpha,n}) \geqq \frac{\sqrt{4\pi\sqrt{3}}}{3} \cdot \sqrt{n_\alpha} \cdot \alpha + \bar{\bar{\varepsilon}}(n,\alpha).$$

Since $L_{\alpha,n} \subset C_{\alpha,n}$ and the length of each of the segments of $C_{\alpha,n}$ is at most α, then $C_{\alpha,n}$ contains at least

$$\frac{\sqrt{4\pi\sqrt{3}}}{3}\sqrt{n_\alpha} + O(\alpha^{-1}) = \frac{\sqrt{4\pi\sqrt{3}}}{3}\sqrt{n_\alpha} + O(\sqrt{n_\alpha})$$

vertices. Thus the theorem is true. □

4. In this section theorems are proved of different form from that in the Introduction. Let S be a class of graphs closed under the subgraph relation. We shall extend the definition of an *$f(n)$-separator theorem* for S to theorems of the following form:

THEOREM B. *Let k be an integer greater than one. There exist constants $\alpha_1 < 1$, $\alpha_2 < 1, \cdots, \alpha_k < 1$ and $\beta > 0$ such that if G is any n-vertex graph in S then the vertices*

of G can be partitioned into $k+1$ sets $A_1, A_2, \cdots, A_{k+1}$ such that no edge joins a vertex in A_i with a vertex in A_j for $1 \leqq i < j \leqq k$, $|A_i| \leqq \alpha_i n$ for $1 \leqq i \leqq k$ and $|A_{k+1}| \leqq \beta f(n)$.

Here we shall prove two theorems of the form of Theorem B. These theorems are useful for some of the applications of the divide-and-conquer strategy. As an example of their use, at the end of this section I prove a theorem of the form of Theorem A with $\alpha = \frac{1}{2}$, which is an improvement of a similar theorem proved in [3].

DEFINITION. Let $G = (V, E)$ be an n-vertex graph and $0 \leqq \gamma \leqq 1$. The sets A, B, C and D are a *regular γ-partition of V* if A, B, C and D partition V, no edge joins a vertex in A with a vertex in B, a vertex in B with a vertex in C or a vertex in C with a vertex in A, $|A| \leqq (1-\gamma)n$, $|B| \leqq (1-\gamma)n$ and $|C| \leqq \gamma n$.

LEMMA 6. *Let $G = (V, E)$ be any n-vertex planar graph and $\frac{1}{2} \leqq \gamma \leqq 1$. Suppose G has a spanning tree of radius r. Then there exists a regular γ-partition of V into four sets A, B, C, D such that D contains no more than $3r+1$ vertices, one the root of the tree.*

Proof. Embed G in the plane. Add a suitable number of additional edges until each face becomes a triangle. Any nontree edge (including the new one) forms a simple cycle with some of the tree edges. The length of this cycle is at most $2r+1$ if it contains the root of the tree, and at most $2r-1$ otherwise. The cycle divides the graph into two parts, the inside and the outside of the cycle.

Let $\delta = 1-\gamma$. Let (x, z) be the nontree edge whose cycle contains (in either its inside or outside region) the minimum (for all cycles and all regions) number of vertices greater than δn. Break ties by choosing the nontree edge whose cycle has the smallest number of faces on the side, where the extremal property occurs. If ties remain, choose arbitrarily.

Suppose without loss of generality that the graph is embedded so that the region with the extremal property is the inside of the (x, z) cycle. Then the outside of the (x, z) cycle contains no more than γn vertices.

Consider the face which has (x, z) as a boundary edge and lies inside the cycle. This face is a triangle and let y be its third vertex. As in [3] determine which of the following cases applies. Figure 3 illustrates the cases.

Case (1). Both (x, y) and (y, z) lie on the cycle. Then the cycle is (x, y, z) and it does not contain any vertices, which is impossible.

Case (2). One of (x, y) and (y, z) (say (x, y)) lies on the cycle. Then (y, z) is a nontree edge and it defines a cycle which contains within it the same vertices as the (x, z) cycle but one face fewer. Then (y, z) would have been chosen in place of (x, z).

Case (3). Neither (x, y) nor (y, z) lies on the cycle.

(3a). Both (x, y) and (y, z) are tree edges. This is impossible since the tree contains no cycles.

(3b). One of (x, y) and (y, z) (say (x, y)) is a tree edge. Then (y, z) is a nontree edge defining a cycle, which contains one vertex (namely y) fewer within it than the (x, z) cycle. Then the inside of the (y, z) cycle contains no more than δn vertices; otherwise (y, z) would have been chosen in place of (x, z). Furthermore, the outside of the (y, z) cycle contains the same vertices as the outside of the (x, z) cycle, which are no more than γn. Let A be the set of vertices inside the (y, z) cycle, let $B = \varnothing$, let C be the set of vertices outside the (y, z) cycle and let D be the set of vertices upon (y, z) cycle. Then the lemma is true.

(3c). Neither (x, y) nor (y, z) is a tree edge. Then each of (x, y) and (y, z) defines a cycle, and each of these cycles contains at least one face fewer within it than the (x, z) cycle. Thus the number of the vertices inside each of the (x, y) and the (y, z) cycles is not greater than δn. Let A be the set of vertices inside the (x, y) cycle, let B be the set of vertices inside the (y, z) cycle, let C be the set of vertices outside the

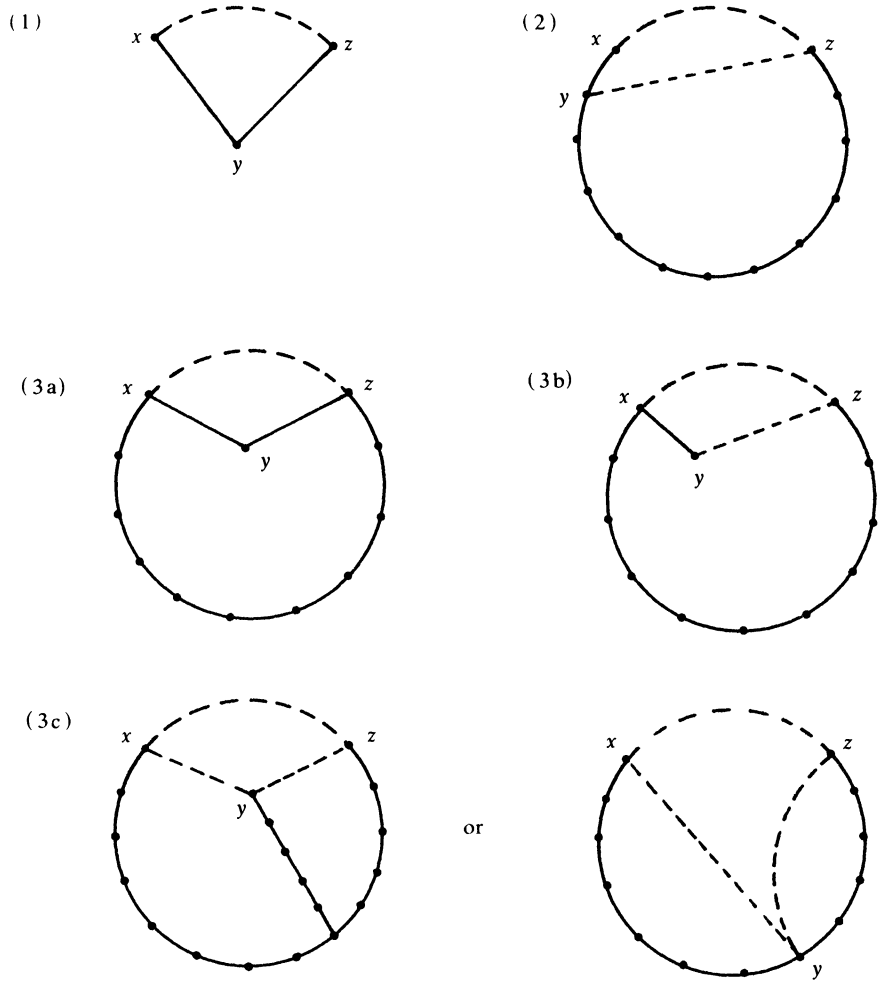


FIG. 3. *Cases for proof of Lemma* 6. *Solid edges are tree edges; dotted edges are nontree edges*

(x, z) cycle, and let D be the set of vertices upon these cycles. The number of the vertices in D is at most $3r+1$ if D contains the root of the tree, and at most $3r-2$ otherwise. □

LEMMA 7. *Let* $G=(V,E)$ *be any* n_0*-vertex graph and* $\frac{1}{2} \leqq \gamma_0 \leqq 1$. *Let* $G^*=(V^*,E^*)$ *be some connected component of* G *containing* $n^* \geqq (1-\gamma_0)n_0$ *vertices. If there exists a constant* k *(depending upon* G^**) such that for each* $\gamma \in [\frac{1}{2}, 1]$ *there exists a regular* γ*-partition of* V^* *into four sets* A^*, B^*, C^*, D^* *such that* $|D^*| \leqq k$, *then there exists a regular* γ_0*-partition of* V *into four sets* A, B, C, D *such that* $|D| \leqq k$.

Proof. Denote $n_1=(1-\gamma_0)n_0$, $n_2=n^*-n_1$, $\gamma^*=(\max\{n_1, n_2\}/n^*)$, $V'=V\backslash V^*$, $n'=n_0-n^*=|V'|$.

Obviously $\gamma^* \in [\frac{1}{2}, 1]$. Let A^*, B^*, C^*, D^* be a regular γ^*-partition of V^* and $|D^*| \leqq k$. Determine the greater number between n_1 and n_2.

If $n_2 \geqq n_1$ let $A=A^*$, $B=B^*$, $C=C^* \cup V'$ and $D=D^*$. Then

$$\gamma^* n^* = \max\{n_1, n_2\} = n_2.$$

Thus

$$\max\{|A|, |B|\} \leqq (1-\gamma^*)n^* = n^* - n_2 = n_1 = (1-\gamma_0)n_0,$$

$$|C| = |C^*| + |V'| \leqq \gamma^* n^* + n' = n_2 + (n_0 - n^*) = n_0 - n_1 = \gamma_0 n_0.$$

If $n_1 > n_2$ let $A = A^*$, $B = C^*$, $C = B^* \cup V'$ and $D = D^*$. Then

$$\gamma^* n^* = \max\{n_1, n_2\} = n_1 = (1-\gamma_0)n_0.$$

Thus

$$|A| \leqq (1-\gamma^*)n^* \leqq \gamma^* n^* = (1-\gamma_0)n_0 \quad \text{since } \gamma^* \geqq \tfrac{1}{2},$$

$$|B| \leqq \gamma^* n^* = (1-\gamma_0)n_0,$$

$$|C| = |B^*| + |V'| \leqq (1-\gamma^*)n^* + n' = n^* + n' - \gamma^* n^* = n_0 - (1-\gamma_0)n_0 = \gamma_0 n_0.$$

Then the lemma is true. □

THEOREM 3. *Let $G = (V, E)$ be any n-vertex planar graph and $\frac{1}{2} \leqq \gamma \leqq 1$. Then there exists a regular γ-partition of V into four sets A, B, C, D such that $|D| \leqq 3\sqrt{2}\sqrt{n}$.*

Proof. Assume G is connected. Partition the vertices of G into levels according to their distance from some vertex v, and let $L(l)$ denote the number of vertices on level l. If r is the maximum distance of any vertex from v, define additional levels -1 and $r+1$ containing no vertices.

Let l_1 be the level for which

$$\sum_{i=0}^{l_1-1} L(i) \leqq (1-\gamma)n \quad \text{and} \quad \sum_{i=0}^{l_1} L(i) > (1-\gamma)n.$$

Let $k = \sum_{i=0}^{l_1} L(i)$. There exist levels $l_0 \leqq l_1$ and $l_2 \geqq l_1 + 1$ such that $L(l_0) + 3(l_1 - l_0) \leqq 3\sqrt{k}$ and $L(l_2) + 3(l_2 - l_1 - 1) \leqq 3\sqrt{n-k}$.

If $\sum_{i=l_0+1}^{l_2-1} L(i) \leqq (1-\gamma)n$ then apparently the theorem is true. Let $\sum_{i=l_0+1}^{l_2-1} L(i) > (1-\gamma)n$. Delete vertices on levels l_0 and l_2 from G and let $G' = (V', E')$ be the resulting graph. By Lemma 6 for the subgraph $G^* = (V^*, E^*)$ of G' induced by the set of vertices on levels $l_0 + 1$ through $l_2 - 1$ in G, and for any $\gamma^* \in [\frac{1}{2}, 1]$ there exists a regular γ^*-partition of V^* into four sets A^*, B^*, C^*, D^* such that $|D^*| \leqq 3(l_2 - l_0 - 1)$.

By Lemma 7 there exists a regular γ-partition of V' into four sets A', B', C', D' such that $|D'| \leqq 3(l_2 - l_0 - 1)$. Thus there exists a regular γ-partition of V into four sets A, B, C, D such that $|D| \leqq L(l_0) + L(l_2) + 3(l_2 - l_0 - 1) \leqq 3(\sqrt{k} + \sqrt{n-k}) \leqq 3\sqrt{2}\sqrt{n}$.

Now suppose G is not connected. Let $G_1, G_2, \cdots, G_k$ be the connected components of G with vertex sets $V_1, V_2, \cdots, V_k$ respectively. If no connected component contains more than $(1-\gamma)n$ vertices, let j be the minimum index such that $\sum_{i=1}^{j} |V_i| > (1-\gamma)n$. Let $A = \cup_{i=1}^{j-1} V_i$, $B = V_j$, $C = \cup_{i=j+1}^{k} V_i$, $D = \varnothing$.

If there exists i, $1 \leqq i \leqq k$ such that $|V_i| > (1-\gamma)n$ then by Lemma 7 there exists a regular γ-partition of V into four sets A, B, C, D such that $|D| \leqq 3\sqrt{2}\sqrt{|V_i|} < 3\sqrt{2}\sqrt{n}$. This completes the proof. □

In the special case when $\gamma = \frac{1}{2}$, using the idea of the proof of Theorem 1, it is possible to reduce the constant $3\sqrt{2} \approx 4.243$ in Theorem 3 to $(3+\sqrt{21})/2 \approx 3.791$.

THEOREM 4. *Let $G = (V, E)$ be any n-vertex planar graph. Then there exists a regular $\frac{1}{2}$-partition of V into four sets A, B, C, D such that $|D| \leqq k\sqrt{n}$, where $k = (3+\sqrt{21})/2$.*

Proof. Assume G is connected. As in Theorem 3, partition the vertices into levels according to their distance from some vertex v and denote with $L(l)$ the number of vertices on level l.

Let l_1 be the minimum index such that $\sum_{l=0}^{l_1} L(l) > n/2$ and let j be the minimum index such that $\sum_{l=l_1-j}^{l_1+j} L(l) > n/2$.

Case 1. There exists i, $0 \leqq i \leqq j$ such that $L(l_1-i)+L(l_1+i) \leqq k\sqrt{n}$. Then the theorem is true.

Case 2. For each i, $0 \leqq i \leqq j$, $L(l_1-i)+L(l_1+i) > k\sqrt{n}$. Let $\beta = (\sum_{l_1-j}^{l_1+j} L(l))/n$. Then $\beta > \frac{1}{2}$. Furthermore

$$\beta n = \sum_{l=l_1-j}^{l_1+j} L(l) > (j+\tfrac{1}{2})k\sqrt{n}.$$

Thus

$$j+\frac{1}{2} < \frac{\beta}{k}\sqrt{n}. \tag{19}$$

Let l' and l'' be levels such that

$$l' \leqq l_1-j-1 < l_1+j+1 \leqq l'',$$

$$L(l')+3(l_1-j-1-l') \leqq 3\sqrt{m}, \tag{20}$$

$$L(l'')+3(l''-(l_1+j+1)) \leqq 3\sqrt{n-m-\beta n}, \tag{21}$$

where $m = \sum_{l=0}^{l_1-j-1} L(l)$. Add together the inequality (19) multiplied by 6, the inequality (20) and the inequality (21). The result is

$$L(l')+L(l'')+3(l''-l'-1) \leqq 3\sqrt{m}+3\sqrt{n-m-\beta n}+6\cdot\frac{\beta}{k}\cdot\sqrt{n}.$$

Since $\sqrt{m}+\sqrt{n-m-\beta n} \leqq \sqrt{2}\sqrt{1-\beta}\,\sqrt{n}$, then

$$L(l')+L(l'')+3(l''-l'-1) \leqq \left(3\sqrt{2}\sqrt{1-\beta}+\frac{6\beta}{k}\right)\sqrt{n}.$$

Let $f(x) = 3\sqrt{2}\sqrt{1-x}+6x/k$, $x \in [\frac{1}{2}, 1]$. Then

$$f'(x) = -\frac{3}{\sqrt{2}\,\sqrt{1-x}}+\frac{6}{k} = \frac{3}{\sqrt{2}\,\sqrt{1-x}\,k}(2\sqrt{2}\,\sqrt{1-x}-k)$$

$$\leqq \frac{3}{\sqrt{2}\cdot\sqrt{1-x}\cdot k}(2-k) < 0 \quad \text{for} \quad x \in [\tfrac{1}{2}, 1).$$

Then $f(\beta) \leqq f(\frac{1}{2})$ and hence

$$3\sqrt{2}\sqrt{1-\beta}+\frac{6\beta}{k} \leqq 3\sqrt{2}\sqrt{\frac{1}{2}}+\frac{3}{k}$$

$$= \frac{6}{3+\sqrt{21}}+3 = \frac{15+3\sqrt{21}}{3+\sqrt{21}} = \frac{3+\sqrt{21}}{2}.$$

Thus $L(l')+L(l'')+3(l''-l'-1) \leqq k\sqrt{n}$.

From this point the proof is the same as in Theorem 3. □

Using both Theorem 3 and Theorem 4 we can now prove the following theorem.

THEOREM 5. *Let $G=(V,E)$ be any n-vertex planar graph. The vertices of G can be partitioned into three sets A, B, C such that no edge joins a vertex in A with a vertex in B,*

$$|A|\leqq\frac{n}{2},\quad |B|\leqq\frac{n}{2},\quad |C|\leqq\left(k+\frac{3\sqrt{2}}{\sqrt{3}-1}\right)\sqrt{n}\quad \textit{where } k=\frac{3+\sqrt{21}}{2}.$$

Proof. We shall define sequences of sets of vertices $\{A_i\}$, $\{B_i\}$, $\{C_i\}$, $\{D_i\}$ such that

(i) A_i, B_i, C_i, D_i partition V.

(ii) No edge joins A_i with B_i, B_i with D_i or D_i with A_i.

(iii) $|A_i|\leqq|B_i|\leqq n/2$.

(iv) $|D_i|\leqq|D_{i-1}|/3$.

Let A^*, B^*, C^*, D^* be a vertex satisfying Theorem 4. Without loss of generality suppose $|A^*|\leqq|B^*|\leqq|C^*|$. Let $A_0=B^*$, $B_0=C^*$, $C_0=D^*$, $D_0=A^*$. Then (i), (ii) and (iii) hold. Furthermore $|D_0|\leqq n/3$ and $|C_0|\leqq k\sqrt{n}$.

Let A_{i-1}, B_{i-1}, C_{i-1} and D_{i-1} be defined and $D_{i-1}\neq\varnothing$. Then $|A_{i-1}|\leqq|B_{i-1}|\leqq n/2$.

Let $n_1=|D_{i-1}|$ and $\gamma=(n/2-|A_{i-1}|)/n_1$. Then

$$(1-\gamma)n_1=n_1-\left(\frac{n}{2}-|A_{i-1}|\right)=(|D_{i-1}|+|A_{i-1}|)-\frac{n}{2}$$

$$=(n-|B_{i-1}|-|C_{i-1}|)-\frac{n}{2}\leqq\frac{n}{2}-|B_{i-1}|.$$

Then

$$(1-\gamma)n_1\leqq\frac{n}{2}-|B_{i-1}|. \tag{22}$$

Furthermore $n/2-|B_{i-1}|\leqq n/2-|A_{i-1}|=\gamma n_1$. Thus

$$(1-\gamma)n_1\leqq\gamma n_1. \tag{23}$$

Let $G^*=(V^*,E^*)$ be the subgraph of G induced by D_{i-1}, and let A^*, B^*, C^*, D^* be a regular γ-partition of V^* satisfying Theorem 3. Then $|D^*|\leqq3\sqrt{2}\sqrt{n_1}$. Without loss of generality (making use of (23)) suppose that $|A^*|\leqq|B^*|\leqq|C^*|$.

Let A_i be the set among $A_{i-1}\cup C^*$ and $B_{i-1}\cup B^*$ with fewer vertices, let B_i be the other set, let $C_i=C_{i-1}\cup D^*$, and let $D_i=A^*$. Since

$$|A_{i-1}\cup C^*|=|A_{i-1}|+|C^*|\leqq|A_{i-1}|+\gamma n_1=\frac{n}{2},$$

$$|B_{i-1}\cup B^*|=|B_{i-1}|+|B^*|\leqq|B_{i-1}|+(1-\gamma)n_1\leqq\frac{n}{2}\quad\text{(by (22))},$$

$$|D_i|=|A^*|\leqq\frac{n_1}{3}=\frac{|D_{i-1}|}{3},$$

then (i)–(iv) hold for A_i, B_i, C_i and D_i.

Let k be the greatest index for which A_k, B_k, C_k and D_k are defined. Then $D_k=\varnothing$. Let $A=A_k$, $B=B_k$, $C=C_k$. By (i), (ii) and (iii) A and B satisfy the requirements of the theorem. By (iv)

$$|C|\leqq k\sqrt{n}+\frac{1}{\sqrt{3}}\sum_{i=0}^{\infty}3\sqrt{2}\sqrt{n}\left(\frac{1}{3}\right)^{i/2}=\left(k+\frac{3\sqrt{2}}{\sqrt{3}(1-\sqrt{\frac{1}{3}})}\right)\sqrt{n}=\left(k+\frac{3\sqrt{2}}{\sqrt{3}-1}\right)\sqrt{n}.$$

□

A similar theorem is proved in [3] with a constant $2\sqrt{2}/(1-\sqrt{\frac{2}{3}}) \approx 15.413$ in the place of $(3+\sqrt{21})/2+3\sqrt{2}/(\sqrt{3}-1) \approx 9.587$.

For all the theorems proved in this paper one can easily construct algorithms finding the appropriate vertex partitions in $O(n)$ time, by modifying the algorithm described in [3].

For some applications of the planar separator theorems see [2].

REFERENCES

[1] A. V. AHO, J. E. HOPCROFT AND J. D. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[2] R. J. LIPTON AND R. E. TARJAN, *Applications of a planar separator theorem*, STAN-CS-77-628, Computer Sci. Dept., Stanford Univ., Stanford, CA, 1977.

[3] R. J. LIPTON AND R. E. TARJAN, *A separator theorem for planar graphs*, STAN-CS-77-627, Computer Sci. Dept., Stanford Univ., Stanford, CA, 1977.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 2, June 1982

0196-5212/82/0302-0010 $01.00/0

COLORING STEINER TRIPLE SYSTEMS*

MARCIA DE BRANDES,† KEVIN T. PHELPS‡ AND VOJTECH RÖDL§

Abstract. In this paper, several results on the chromatic number of Steiner triple systems are established. A Steiner triple system is a simple 3-uniform hypergraph in which every pair of vertices is connected by exactly one 3-edge. Among other things, we prove that for any $k \geqq 3$ there exists an n_k such that for all admissible $v \geqq n_k$ there exists a k-chromatic Steiner triple systems of order v. In addition we prove that for all $v \geqq 49$ there exists a 4-chromatic Steiner triple system of order v. An estimate of n_k is also established, namely, $c_1 k^2 \log k > n_k > c_2 k^2$.

1. Introduction. A Steiner triple system of order v (briefly STS (v)) is a pair (S, B), where S is a v-set and B is a collection of 3-subsets of S called triples, such that every 2-subset of S is contained in exactly one triple of B. Such a triple system can also be considered as a special 3-uniform hypergraph; in this light questions involving chromatic number and colorings may appear more natural. The definitions are the same as for hypergraphs: a (proper) k-coloring of a Steiner triple system (S, B) is a partition of S into k color classes such that no triple in B is monochromatic (that is, is properly contained in any color class). If an STS can be k-colored but not $(k-1)$-colored, then it is said to be k-chromatic.

The chromatic number of Steiner triple systems has previously been investigated by Rosa [11], [12]. Among other things he established that there exists a 3-chromatic STS of all admissible orders (excluding the trivial systems on 1 and 3 elements, respectively). He also gave some constructions for 4-chromatic STSs. In this paper we will show that there exists a 4-chromatic STS (v) for all $v \geqq 25$, $v \equiv 1$ or 3 (mod 6), except possibly for $v = 39$, 43 and 45. As a part of the proof of this, we introduce two color preserving recursive constructions that work for STSs with arbitrary chromatic number $k \geqq 4$.

A partial triple system differs from a (complete) Steiner triple system in that any 2-subset is contained in *at most* one triple of the system. It is not difficult to see that a 3-uniform hypergraph without short cycles is in fact a partial triple system. It has been shown that for any k there exists such a hypergraph which is k-chromatic [6], [7]. It is then immediate [12] that the previous result—in conjunction with a result by Treash [15] that every partial triple system can be embedded in a (complete) STS—shows that STS can have arbitrarily large chromatic number. Unfortunately, there is no known embedding that will necessarily preserve the chromatic number of the partial triple system (in a sense, none could: a partial triple system can be 2-chromatic whereas a (nontrivial) STS must be at least 3-chromatic). In the next section of this paper, we get around this difficulty and manage to prove that for any $k \geqq 3$ there exists an n_k such that for all admissible orders v, $v \geqq n_k$, there exists a k-chromatic STS (v).

2. k-chromatic triple systems. As stated already, there exists a k-chromatic partial triple system on u_k elements (see [6], [7]). In both [6] and [7], a more general question is studied and the proofs of the above result do not give a reasonable upper bound for u_k (e.g., [7, Thm. 1] gives $u_k < ck^4$). For this reason we present the following.

* Received by the editors January 19, 1981.
† Department of Mathematical Sciences, McMaster University, Hamilton, Ontario, Canada L8S 4K1.
‡ School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.
§ FJFI ČVUT, Husova 5, Praha 1, Czechoslovakia.

LEMMA 2.1.

$$c_1 k^2 \log k > u_k > c_2 k^2.$$

The lower bound follows immediately from [7, Thm. 2] (with the value $c_2 \sim \frac{1}{6}$). We will prove the upper bound by the probabilistic method. First we need to introduce some notions. If Ω is a probability space and $A_1, A_2, \cdots, A_n$ events, denote by Γ a graph with vertex set $\{1, 2, \cdots, n\}$, where $\{i, j\} \notin \Gamma$ if and only if A_i and A_j are mutually independent. The key to our proof is the following theorem of Spencer [14], which in turn is a consequence of a theorem of Lovász (cf. [7], [14]).

THEOREM 2.1 [14]. *Let $A_1, A_2, \cdots, A_n$ be events in the probability space with dependence graph Γ. If there exist positive $y_1, y_2, \cdots, y_n$ with $y_i P(A_i) < 1$ such that*

$$\log y_i > \sum_{\{i,j\} \in \Gamma} y_j P(A_j),$$

then $P(\bigwedge \bar{A}_i) > 0$.

This theorem is used in [14] to give a simple proof of the lower bound for the Ramsey number $R(3, t) \geqq ct^2/\log^2 t$ (cf. [5]). As the method we will use here is a modification of the above proof, we will preserve the same notation as in [14].

Proof of Lemma 2.1 (*upper bound*). Let V be a set with $m = c_1 k^2 \log k$ elements. Consider a random 3-uniform hypergraph G with vertex V where the triples are chosen independently, each with probability $p = c/m$. If L is a set of 4 vertices, let A_L be the event that $|G \cap [L]^3| \geqq 2$. If K is a set of m/k vertices, let B_K be the event that $G \cap [K]^3 = \varnothing$. Clearly, if

$$P\Big(\bigwedge_{L \in [V]^4} \bar{A}_L \wedge \bigwedge_{K \in [V]^{m/k}} \bar{B}_K \Big) > 0,$$

then $u_k \leqq m$. Let Ω be a space with events A_L, B_K, $L \in [V]^4$, $K \in [V]^{m/k}$; thus two vertices of Γ corresponding to events A_L, B_K will be joined if $|L \cap K| \geqq 3$. Similarly, vertices corresponding to $A_L, A_{L'}$ (or $B_K, B_{K'}$) will be joined if $|L \cap L'| \geqq 3 (|K \cap K'| \geqq 3$, respectively). As in Spencer [14], one can define N_{AA}, N_{AB}, N_{BA}, N_{BB} to be the number of vertices, N_{XY}, in Γ of type Y adjacent to a vertex of type X. We also associate with each event $A_L (B_K)$ some $y_L = y$ ($z_K = z$). Then it suffices to prove that there exist y, z such that

$$\begin{aligned} & yP(A_L) < 1, \qquad zP(B_K) > 1, \\ (*) \qquad & \log y > yP(A_L) N_{AA} + zP(B_K) N_{AB}, \\ & \log z > yP(A_L) N_{BA} + zP(B_K) N_{BB}. \end{aligned}$$

Set $t = m/k$. Clearly

$$P(A_L) \leqq \binom{4}{2} p^2 = \frac{6c^2}{m^2}, \qquad P(B_K) = (1-p)^{\binom{t}{3}} \sim \exp\left(-\frac{pt^3}{6}\right),$$

$$N_{AB} \leqq \binom{m}{t} < \left(\frac{me}{t}\right)^t < m^{t/2}, \qquad N_{BB} < m^{t/2},$$

$$N_{AA} \leqq 4(m-4) < 4m, \qquad N_{BA} < \binom{t}{3}(m-t) + \binom{t}{4} < \frac{t^3 m}{6}.$$

Set $y = 1 + 1/t$, $z = \exp(\xi m^{1/2} \log^{3/2} m)$. Then (*) becomes

$$\frac{(1+1/t)6c^2}{m^2} < 1, \qquad \xi - \frac{cd^{3/2}}{6} < 0, \quad \text{where } d = \frac{c_1}{2}(1+o(1)),$$

$$\log\left(1+\frac{1}{t}\right) > \frac{(1+1/t)24c^2}{m} + \exp\left(\left(\xi - \frac{cd^{3/2}}{6} + \frac{\sqrt{d}}{2}\right) m^{1/2} \log^{3/2} m\right),$$

$$\xi m^{1/2} \log^{3/2} m > \left(1+\frac{1}{t}\right) c^2 d^{3/2} m^{1/2} \log^{3/2} m + \exp\left(\left(\xi - \frac{cd^{3/2}}{6} + \frac{\sqrt{d}}{2}\right) m^{1/2} \log^{3/2} m\right).$$

It follows by elementary calculations that the above conditions are satisfied for k sufficiently large if

$$(**) \qquad \xi + \frac{\sqrt{d}}{2} - \frac{cd^{3/2}}{6} < 0, \qquad c^2 d^{3/2} - \xi < 0.$$

We want to find minimum t, and hence d, such that (**) is satisfied. It follows from elementary analysis that the corresponding values are $d = 72(1+o(1))$ and $c = 1/12(1+o(1))$ $\xi = 18(1+o(1))$, and thus

$$u_k \leqq 144(1+o(1))k^2 \log k.$$

Lindner [9] has given a small and simple embedding of partial triple systems. Again, the embedding will not necessarily preserve the chromatic number of the partial system. However, by using Lindner's approach we can construct k-chromatic STS for all sufficiently large v, $v \equiv 1$ or 3 (mod 6). First we need to present two constructions for STSs, one due to Bose [2], the other originally due to Skolem [13]. The presentation of these constructions is taken from the excellent survey article [9].

Bose's construction. Let $(Q, \circ)$ be an indempotent commutative quasigroup of order $2u+1$, i.e., a quasigroup satisfying the identities $x \circ x = x$, and $x \circ y = y \circ x$. Let $Q = \{1, 2, 3, \cdots, 2u+1\}$ and $S = Q \times \{1, 2, 3\}$. Define a collection of triples t of S as follows:

(1) $\{(x, 1), (x, 2), (x, 3)\} \in t$ for every $x \in Q$;

(2) if $x \neq y$, then $\{(x, 1), (y, 1), (x \circ y, 2)\}$, $\{(x, 2), (y, 2), (x \circ y, 3)\}$ and $\{(x, 3), (y, 3), (x \circ y, 1)\} \in t$.

It is a routine matter to see that (S, t) is an STS $(6u+3)$.

Skolem's construction. A quasigroup $(Q, \circ)$ with $Q = \{1, 2, 3, \cdots, 2v\}$ satisfying

$$x \circ x = \begin{cases} x & \text{if } x \leqq v, \\ x - v & \text{if } x > v \end{cases}$$

is called a halfidempotent quasigroup [9]. If the quasigroup is both halfidempotent and commutative, then its order must be even. Now let $(Q, \circ)$ be a halfidempotent commutative quasigroup of order $2v$ and set $S = Q \times \{1, 2, 3\} \cup \{\infty\}$. Define a collection of triples t of S as follows:

(1) $\{(x, 1), (x, 2), (x, 3)\} \in t$ for every $x \leqq v$, $x \in Q$;

(2) for each $x > v$, the three triples $\{\infty, (x, 1), (x-v, 2)\}\{\infty, (x, 2), (x-v, 3)\}$ and $\{\infty, (x, 3), (x-v, 1)\} \in t$;

(3) if $x \neq y$, then $\{(x, 1), (y, 1), (x \circ y, 2)\}$, $\{(x, 2), (y, 2), (x \circ y, 3)\}$ and $\{(x, 3), (y, 3), (x \circ y, 1)\} \in t$.

Again (S, t) is an STS $(6v+1)$.

Let us remark that in Bose's construction subquasigroups will always produce subsystems in the resulting STS. A similar thing holds for Skolem's construction as well.

As is well known, one can construct a (partial) idempotent commutative quasigroup $(Q, \circ)$ from a (partial) triple system: define $x \circ x = x$ and $x \circ y = y \circ x = z$ if and only if the third point of the triple containing the pair $\{x, y\}$ is z. If the pair $\{x, y\}$ is not contained in any triple of the partial triple system then the operation $\circ$ is not defined for that product. Next we state two results which establish that a partial idempotent commutative quasigroup can be embedded in both idempotent and half-idempotent commutative quasigroups.

LEMMA 2.2 (A. Cruse [3]). *A partial idempotent commutative quasigroup of order n can be embedded in an idempotent commutative quasigroup of order t for every odd* $t \geqq 2n+1$ *and in a commutative quasigroup of order t for every* $t \geqq 2n$.

LEMMA 2.3 (C. C. Lindner [9]). *A partial idempotent commutative quasigroup of order n can be embedded in a halfidempotent commutative quasigroup of order* $2t$ *for all* $t \geqq n$.

Now for the main theorem of this section.

THEOREM 2.2. *For all* $k \geqq 3$ *there exists an* n_k *such that for every* $v \equiv 1$ *or* $3 \bmod 6$, $v \geqq n_k$, *there exists a k-chromatic* STS (v).

Proof. By Lemma 2.1 we know that there exists a k-chromatic partial triple system of order u_k. From this partial triple system one can construct a partial idempotent quasigroup and then embed it in an idempotent (or halfidempotent) commutative quasigroup of order $2t+1$ (or $2t$, respectively) for every $t \geqq u_k$. Apply Bose's construction to this idempotent commutative quasigroup and Skolem's construction to this halfidempotent commutative quasigroup, as the case may be. This gives us Steiner triple systems of orders $6t+3$ and $6t+1$ for all $t \geqq u_k$. It is immediate from these constructions that the resulting STSs are 3-chromatic. Moreover, for every triple $\{x, y, z\}$ in the original partial triple system we have a subsystem of order 9 on the set $\{x, y, z\} \times \{1, 2, 3\}$ in the resulting STS. Steiner triple systems have the replacement property, i.e., one can remove or "unplug" the triples of any subsystem and replace them with the triples of any other subsystem defined on the same subset. The resulting collection of triples is still an STS. In particular, for each triple b in the original partial triple system we can replace the subsystem on $b \times \{1, 2, 3\}$ with a system of order 9 which contains the following blocks (among others): for each $x \in b$, $\{(x, 1), (x, 2), (x, 3)\}$ is in the subsystem and $b \times \{1\}$ is also in this subsystem. This gives us a new collection of triples. We can repeat this procedure for each of the original triples giving a sequence of STSs; the final STS in this sequence contains a copy of the original partial triple system embedded in it and, hence, must have chromatic number *at least k*. We claim that a single replacement of a subsystem of order 9 as described above increases the chromatic number by at most one, and hence, one of the Steiner triple systems in this sequence must be k-chromatic.

Suppose the current STS in our sequence is i-chromatic. Suppose we construct the next triple system in this sequence and consider any (proper) i-coloring of the previous triple system. It may still be a proper coloring of the new triple system. However, if it is not then the only monochromatic triples must be contained in the new subsystem of order 9. An STS (9) is 3-chromatic, and it is a trivial exercise to show that in any 3-coloring of it no color class can have more than 4 elements. Again it is trivial to see that one can choose one element from each monochromatic triple (there are at most 3) so that the resulting set contains no triple of the subsystem of order 9 (and, hence, can contain no triple of the system either). Assigning $(i+1)$st color to this set obviously gives a proper $(i+1)$-coloring. Hence the new triple system has chromatic number i or $i+1$. This completes the proof of Theorem 2.4.

3. 4-chromatic Steiner triple systems. In § 2 we established the existence of a k-chromatic STS (v) for all sufficiently large orders v. To facilitate further discussion, let us introduce some notation: n_k will denote the smallest admissible integer such that there exists a k-chromatic STS (v) for all admissible $v \geqq n_k$. The results of the previous section give an upper bound on $n_k \leqq ck^2 \log k$, where $c = 864(1+o(1))$. It was shown previously [12] that $n_3 = 7$. It seems that the only other value of k for which it remains practical to determine n_k—at least at present—is $k = 4$. In this section we show that $n_4 \leqq 49$ (although we believe that in fact $n_4 = 25$).

LEMMA 3.1. *There exists a* 4*-chromatic* STS (v) *for* $v = 25$, 27, 33 *and* 37.

Proof. The triple systems below have integers $1, 2, \cdots, v$ as elements and are all cyclic; i.e., they have as an automorphism the map $i \to i+1 (\mathrm{mod}\ v)$. For each v the base triples of the STS along with a 4-coloring are given. A computer was used to establish that no 3-coloring exists for any of these systems.

$v = 25$: Base triples: {1, 2, 4}, {1, 5, 24}, {1, 6, 12}, {1, 8, 18}.

4-coloring: {1, 2, 3, 6, 7, 8, 11}
{5, 9, 10, 13, 14, 15, 19}
{12, 16, 17, 18, 21, 22}
{4, 20, 23, 24, 25}.

$v = 27$: Base triples: {1, 2, 4}, {1, 5, 12}, {1, 6, 18}, {1, 7, 15},
{1, 10, 19}.

4-coloring: {1, 2, 3, 6, 7, 8, 11, 12, 22, 25}
{10, 13, 14, 15, 18, 19, 20, 23, 24}
{4, 5, 9, 17, 26}
{16, 21, 27}.

$v = 33$: Base triples: {1, 2, 4}, {1, 5, 15}, {1, 6, 14}, {1, 7, 19},
{1, 8, 17}, {1, 12, 23}.

4-coloring: {1, 2, 3, 6, 7, 8, 12, 13, 30}
{14, 15, 16, 19, 20, 21, 24, 25, 26, 31}
{4, 5, 11, 27, 28, 29, 32, 33}
{9, 10, 17, 18, 22, 23}.

$v = 37$: Base triples: {1, 2, 4}, {1, 5, 15}, {1, 6, 14}, {1, 7, 22},
{1, 8, 20}, {1, 10, 21}.

4-coloring: {1, 2, 3, 6, 7, 8, 11, 12, 13, 18, 32}
{14, 15, 16, 19, 20, 21, 24, 25, 26, 37}
{4, 9, 23, 29, 30, 31, 34, 35, 36}
{5, 10, 17, 22, 27, 28, 33}.

It is our feeling that we have examples of 4-chromatic Steiner triple systems of orders 39, 43 and 45 as well. However, to prove that our systems are not 3-chromatic is currently beyond our resources—computer and otherwise.

Next we present two simple recursive constructions (Lemmas 3.2–3.4) preserving the chromatic number of STSs. Although in this paper we do not make use of the second construction, we would like to point out that Lemmas 3.2–3.4 could be used for an alternative proof of Theorem 2.2 (say, if Skolem's construction were not available). We also feel that besides being of interest on their own these constructions may prove useful when considering STSs with chromatic number $k \geqq 5$.

In what follows let (S, B) be an STS (v), where $S = \{a_1, a_2, \cdots, a_v\}$. (To avoid trivial cases, assume $v \geqq 7$.) Whenever C is a k-coloring of (S, B) with $k \geqq 4$, there must be some $k-2$ colors such that at least $(v+1)/2$ elements of S have these colors. We will always assume these $k-2$ colors to be all colors except black and white and assume that these (at least) $(v+1)/2$ elements that are colored other than black or white are $a_1, a_2, \cdots, a_{(v+1)/2}$.

If one can find $k-2$ colors such that at least $(v+3)/2$ elements of S have these colors (in which case we will again assume these colors to be all colors except black and white and the corresponding (at least) $(v+3)/2$ elements that are colored other than black or white to be $a_1, a_2, \cdots, a_{(v+3)/2}$), then the coloring will be called *biased.* Observe that when $k \geqq 5$, every k-coloring is biased. However, there exist STSs with an unbiased 4-coloring. On the other hand, we do not know an example of a 4-chromatic STS without a biased 4-coloring.

LEMMA 3.2. *If there exists a k-chromatic* STS (v), *then there exists a k-chromatic* STS $(2v+1)$.

Proof. The lemma is obviously true for $k = 3$, so we may assume $k \geqq 4$. Let (S, B) be a k-chromatic STS (v), where $S = \{a_1, a_2, \cdots, a_v\}$, and let C be a k-coloring of (S, B). Put $v+1 = 2n$, and let T be a set such that $|T| = 2n$, $S \cap T = \varnothing$. Let $T = T_1 \cup T_2$ be any partition of T with $|T_1| = |T_2| = n$. Distinguish now two cases:

Case 1. $n \equiv 0 \pmod 2$. Let (T, F) be any $OF(K_{2n})$ having sub-$OF(K_n)$ of index $2(T_i, F^i)$, where $F = \{F_1, F_2, \cdots, F_{2n-1}\}$, $F^i = \{F_1^i, F_2^i, \cdots F_{n-1}^i\}$, $i = 1, 2$, and let $F_j = F_j^1 \cup F_j^2$ for $j = 1, 2, \cdots, n-1$.

Case 2. $n \equiv 1 \pmod 2$. In this case let (T, F) be an $OF(K_{2n})$ with the following properties: $F = \{F_1, F_2, \cdots, F_{2n-1}\}$, $F_j = F_j^1 \cup F_j^2 \cup \{x_j, \alpha(x_j)\}$ for $j = 1, 2, \cdots, n$, where (T_i, F^i), $i = 1, 2$, $F^i = \{F_1^i, F_2^i, \cdots, F_n^i\}$, is a near-$OF(K_n)$, $T_1 = \{x_1, x_2, \cdots, x_n\}$ and $\alpha : T_1 \to T_2$ is any bijection. Such $OF(K_{2n})$ is well known to exist (see, e.g., [1]).

In either case, put $S^* = S \cup T$ and $B^* = B \cup D$, where

$$D = \{\{a_1, x, y\} \mid \{x, y\} \in F_i,\ i = 1, 2, \cdots, 2n-1\}.$$

Then (S^*, B^*) is an STS $(2v+1)$ (cf. [10]). Moreover, (S^*, B^*) is k-chromatic: color the elements of T_1 black, those of T_2 white, and let the elements of S have the same color as in the coloring C. There are no monochromatic triples: this is obviously true for triples of B. If $\{a_i, x, y\}$ is a triple of D with $i \in \{1, 2, \cdots, n\}$, then a_i is colored by one of the $k-2$ colors other than black or white while x, y $(\in T)$ can be only black or white. If $i \in \{n+1, n+2, \cdots, 2n-1\}$ then one of x, y is black and the other is white.

LEMMA 3.3. *Let* $v \equiv 1$ *or* 9 $\pmod{12}$. *If there exists a k-chromatic* STS (v), *then there exists a k-chromatic* STS $(2v+7)$.

Proof. We may again assume $k \geqq 4$, and let (S, B) be a k-chromatic STS (v) with $S = \{a_1, a_2, \cdots, a_v\}$, with C a k-coloring of (S, B). Put $v+7 = 2m$; then m is even since $v \equiv 1$ or 9 $\pmod{12}$. Let $X = \{x_1, x_2, \cdots, x_m\}$, $Y = \{y_1, y_2, \cdots, y_m\}$, $X \cap Y = \varnothing$, $T = X \cup Y$, $T \cap S = \varnothing$. Let (X, F), $F = \{F_1, F_2, \cdots, F_{m-1}\}$ be an $OF(K_m)$ containing

two 1-factors (let these be, w.l.o.g., F_{m-2} and F_{m-1}) whose union is a Hamiltonian circuit (let, again w.l.o.g., this Hamiltonian circuit be $F_{m-2} \cup F_{m-1} = (x_1\, x_2 \cdots x_m\, x_1)$). Such an $OF(K_m)$ is well known to exist (cf., e.g., [10]).

Let

$$C = \{\{y_i, x_{i+3}, x_{i+4}\}, \{y_i, y_{i+1}, x_{i+2}\} \mid i = 1, 2, \cdots, m\},$$

$$D = \{\{a_i, x_p, x_q\}, \{a_i, y_p, y_q\} \mid \{x_p, x_q\} \in F_i,\ i = 1, 2, \cdots, m-3\},$$

$$E = \{\{a_{m-2+k}, x_j, y_{j+k}\} \mid j = 1, 2, \cdots, m;\ k = 0, 1, \cdots, m-5\}$$

(the subscripts of x's and y's reduced modulo m to the range $\{1, 2, \cdots, m\}$ whenever necessary).

Put $S^* = S \cup T$, $B^* = B \cup C \cup D \cup E$. It is easily verified that (S^*, B^*) is an STS $(2v+7)$. To show that (S^*, B^*) is k-chromatic, color the elements of X black, those of Y white and those of S as in the coloring C. There are no monochromatic triples in B^*. This is obvious for triples of B and also those of C and E as the latter two contain only triples with at least one black and at least one white element. On the other hand, no element a_i with $i \in \{1, 2, \cdots, m-3\}$ is colored black or white; thus no triple of D can be monochromatic.

LEMMA 3.4. *Let $v \equiv 3$ or 7 (mod 12). If there exists a k-chromatic* STS (v) *with a biased k-coloring, then there exists a k-chromatic* STS $(2v+7)$.

Proof. We may assume again $k \geqq 4$. Let (S, B) be a k-chromatic STS (v) with $S = \{a_1, a_2, \cdots, a_v\}$, and let C be a biased k-coloring of C. Put $v+7 = 2m$ (then $m \equiv 1$ (mod 2)). Let $X = \{x_1, x_2, \cdots, x_m\}$, $Y = \{y_1, y_2, \cdots, y_m\}$, $X \cap Y = \varnothing$, $T = X \cup Y$, $T \cap S = \varnothing$. Let (X, F) be a near-$OF(K_m)$, $F = \{F_1, F_2, \cdots, F_m\}$ containing two near-1-factors (let these be, w.l.o.g., F_{m-1}, F_m) whose union is a hamiltonian path (let, again w.l.o.g., this hamiltonian path be $F_{m-1} \cup F_m = (x_1\, x_2 \cdots x_m)$). Such a near-$OF(K_m)$ is easily seen to exist: it can be obtained, e.g., from a 1-factorization GK_{m+1} (cf. [1], [10]) by omitting any one vertex. Assume further w.l.o.g. that the edge $\{x_1, x_m\}$ belongs to the factor F_{m-2}.

Let now

$$C = \{\{y_i, x_{i+3}, x_{i+4}\}, \{y_i, y_{i+1}, x_{i+2}\} \mid i = 1, 2, \cdots, m\},$$

$$D = \{\{a_i, x_p, x_q\}, \{a_i, y_p, y_q\} \mid \{x_p, x_q\} \in F_i,\ i = 1, 2, \cdots, m-3\},$$

$$D' = \{\{a_i, x_{j(i)}, y_{j(i)}\} \mid i = 1, 2, \cdots, m-2,\ x_{j(i)} \text{ the isolated vertex of } F_i\},$$

$$D'' = \{\{a_{m-2}, x_p, x_q\}, \{a_{m-2}, y_p, y_q\} \mid \{x_p, x_q\} \in F_{m-2} \setminus \{x_1, x_m\}\}$$
$$\cup \{\{a_{m-2}, x_1, y_1\}, \{a_{m-2}, x_m, y_m\}\},$$

$$E = \{\{a_{m-2+k}, x_j, y_{j+k}\} \mid j = 1, 2, \cdots, m;\ k = 1, 2, \cdots, m-5\}$$

(the subscripts of x's and y's reduced modulo m to the range $\{1, 2, \cdots, m\}$ whenever necessary).

Put $S^* = S \cup T$, $B^* = B \cup C \cup D \cup D' \cup E$. It is again easily verified that (S^*, B^*) is an STS $(2v+7)$. Moreover, (S^*, B^*) is k-chromatic: if one colors elements of X black, those of Y white and those of S as in the coloring of (S, B), then there are no monochromatic triples of B^*. This is obvious for triples of B, C, D' and E. Since $m-2 = (v+3)/2$ and the coloring C of (S, B) is biased, neither of the elements $a_1, a_2, \cdots, a_{m-2}$ is black or white; thus, no triple of D or D' can be monochromatic, either.

Remark. Observe that whenever one applies Lemma 3.2, 3.3 or 3.4 the resulting k-coloring of STS (u), where $u = 2v + 1$ or $u = 2v + 7$ and $k \geqq 4$, is automatically biased, and so Lemma 3.4 can certainly be applied repeatedly even when $k = 4$.

LEMMA 3.5. *There exists a* 4*-chromatic* STS (v) *for* $v \geqq 49$.

Proof. We utilize the Bose and Skolem constructions of Steiner triple systems which are presented in the previous section.

Case 1. $v \equiv 1 \bmod 6$. Then $v = 3u + 1$, where $u \equiv 0 \bmod 2$ and $u \geqq 18$. For each of these values of u, there exists a halfidempotent commutative latin square of order u which contains a halfidempotent commutative subsquare of order 8 [9]. Skolem's construction, as described in [9], applied to such a latin square produces a 3-chromatic STS $(3u + 1)$ with color classes of size u, u and $u + 1$. Moreover, such STS will possess a 3-chromatic sub-STS (25) with color classes of size 8, 8, 9. Unplug this subsystem of order 25, and replace it with a 4-chromatic STS (25) in such a way that 3 of the 4 color classes are subsets of the 3 existing color classes (of size 8, 8, 9). Clearly, the result is a 4-chromatic STS (v) for $v = 3u + 1$.

Case 2. $v \equiv 3 \bmod 6$. Then $v = 3u$ for $u \equiv 1 \bmod 2$ and $u \geqq 19$. There exists a commutative idempotent latin square of order u, $u \equiv 1 \bmod 2$, $u \geqq 19$ which contains a commutative idempotent subsquare of order 9. Using such a latin square in Bose's construction of an STS $(3u)$ gives a 3-chromatic STS with a (3-chromatic) subsystem of order 27. As before, we replace this subsystem with a 4-chromatic STS (27) so that 3 of its color classes are subsets of the original color classes. Obviously, the result is a 4-chromatic STS (v) for $v = 3u$, where $v \geqq 57$.

Case 3. $v = 49$ or 51. For $v = 49$ see [12]. For $v = 51$ set $v = 2u + 1$, where $u = 25$. Apply Lemma 3.2 to the 4-chromatic STS (25) presented in Lemma 3.1.

The main theorem of this section now follows from the previously established lemmas.

THEOREM 3.6. *There exists a* 4*-chromatic* STS (v) *for all* $v \geqq 25$, $v \equiv 1$ *or* 3 (mod 6), *except possibly* $v = 39$, 43 *or* 45.

Thus, $n_4 \leqq 49$ as claimed at the beginning of this section. We conjecture, however, that $n_4 = 25$.

4. Conclusion and open problems. In our search for 4-chromatic Steiner triple systems of small orders, we discovered a uniquely colorable 3-chromatic STS of order 33. A natural question arises: do there exist uniquely colorable k-chromatic Steiner triple systems for all k?

It was shown recently [5] that there exists a polynomial algorithm for deciding whether a Steiner quadruple system is 2-chromatic. How difficult is it to decide whether an STS is k-chromatic? Experience seems to indicate that it is difficult to decide even whether an STS is 3-chromatic.

There are many further questions that can be asked; let us mention just one more problem: Let $C(v) = \{k$: there exists a k-chromatic STS $(v)\}$. Is $C(v)$ an interval? We expect the answer to this to be "yes".

REFERENCES

[1] B. A. ANDERSON, *Symmetry groups of some perfect* 1*-factorizations of complete graphs*, Discrete Math., 18 (1977), pp. 227–254.

[2] R. C. BOSE, *On the construction of balanced incomplete block designs*, Ann. Eugenics, 9 (1939), pp. 353–399.

[3] A. CRUSE, *On embedding incomplete symmetric latin squares*, J. Combinat. Theory (A), 18 (1975), pp. 349–351.

[4] C. J. COLBOURN, M. J. COLBOURN, K. T. PHELPS AND V. RÖDL, *Coloring Steiner quadruple systems*, Discr. Appl. Math., to appear.
[5] P. ERDÖS, *Graph theory and probability* II, Canad. J. Math., 13 (1961), pp. 346–352.
[6] P. ERDÖS AND A. HAJNAL, *On the chromatic number of graphs and set systems*, Acta Math. Acad. Sci. Hungar., 17 (1966), pp. 61–99.
[7] P. ERDÖS AND L. LOVÁSZ, *Problems and results on* 3*-chromatic hypergraphs and related questions*, Infinite and Finite Sets, Proc. Conf. Keszthely 1973, Colloq. Math. Soc. J. Bolyai 10, pp. 609–617.
[8] C. C. LINDNER, *A partial Steiner triple system of order n can be embedded in a Steiner triple system of order* $6n+3$, J. Combinat. Theory (A), 18 (1975), pp. 349–351.
[9] ———, *A survey of embedding theorems for Steiner systems*, Topics on Steiner Systems, Ann. Discrete Math., 7 (1980), pp. 175–202.
[10] C. C. LINDNER, E. MENDELSOHN AND A. ROSA, *On the number of* 1*-factorizations of the complete graph*, J. Combinat. Theory (B), 20 (1976), pp. 265–282.
[11] A. ROSA, *On the chromatic number of Steiner triple systems*, Combinatorial Structures and Their Applications (Proc. Conf. Calgary 1969), Gordon and Breach, New York, 1970, pp. 369–371.
[12] ——— *Steiner triple systems and their chromatic number*, Acta Fac. Rerum Natur. Univ. Comen. Math., 24 (1970), pp. 159–174.
[13] T. SKOLEM, *Some remarks on the triple systems of Steiner*, Math. Scand., 6 (1968), pp. 273–280.
[14] J. SPENCER, *Asymptotic lower bounds for Ramsey functions*, Discrete Math., 20 (1977), pp. 69–77.
[15] C. A. TREASH, *The completion of finite incomplete Steiner triple systems with applications to loop theory*, J. Combinat. Theory (A), 10 (1971), pp. 259–265.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 2, June 1982

0196-5212/82/0302-0011 $01.00/0

ON THE *LU* FACTORIZATION OF *M*-MATRICES: CARDINALITY OF THE SET $\mathscr{P}_n^g(A)$

R. S. VARGA† AND D.-Y. CAI‡

Abstract. An $n \times n$ M-matrix A is said to admit an LU factorization into $n \times n$ M-matrices if A can be expressed as $A = LU$ where L is an $n \times n$ lower triangular M-matrix and where U is an upper triangular M-matrix. Then, for any given $n \times n$ M-matrix A, let $\mathscr{P}_n^g(A)$ denote the set of all $n \times n$ permutation matrices P such that PAP^T admits an LU factorization into M-matrices with nonsingular L. Our aim here is to determine upper and lower bounds for $|\mathscr{P}_n^g(A)|$, the cardinality of the set $\mathscr{P}_n^g(A)$. This is done in Theorem 4, while in Theorem 2, $|\mathscr{P}_n^g(A)|$ is precisely determined for a special class of $n \times n$ M-matrices.

1. Introduction. If the spectrum, $\sigma(B)$, of an $n \times n$ complex matrix B is defined as

$$\sigma(B) := \{\lambda \in \mathbb{C}: \det[\lambda I - B] = 0\}, \tag{1.1}$$

then an $n \times n$ real matrix $A = [a_{i,j}]$ is said to be an *M-matrix* if

$$a_{i,j} \leqq 0 \quad \text{for all } i \neq j, \quad 1 \leqq i, j \leqq n, \tag{1.2}$$

and if

$$\operatorname{Re} \lambda \geqq 0 \quad \text{for all } \lambda \in \sigma(A). \tag{1.3}$$

It may be somewhat surprising to learn that, despite such a simple definition, the theory and applications of M-matrices form one of the *major* building-blocks of numerical linear algebra (cf. [1] and [6]). Moreover, the applications of M-matrices extend beyond numerical linear algebra to Markov chains, input-output economic models, dynamical systems, mathematical programming, and the compartmental analysis of ecological systems (cf. [1] and [3]).

Such an application as above can give rise to a large sparse system of linear equations whose associated coefficient matrix A is an $n \times n$ M-matrix. For direct methods, comparable to the Gaussian elimination method for solving this system of linear equations, it is of practical interest to know if the associated M-matrix A can be factored as $A = L \cdot U$, where L is an $n \times n$ lower triangular M-matrix, and where U is an upper triangular M-matrix. More precisely, as in [7], an $n \times n$ M-matrix A is said to *admit an LU factorization into $n \times n$ M-matrices*, if A can be expressed as

$$A = LU, \tag{1.4}$$

where L is an $n \times n$ lower triangular M-matrix and where U is an $n \times n$ upper triangular M-matrix. As shown in 1962 by Fiedler and Pták [2], any nonsingular M-matrix admits such an LU factorization into M-matrices, with both L and U nonsingular. In 1977, Kuo [5] extended this result by showing that any $n \times n$ irreducible M-matrix (singular or not) admits an LU factorization (1.4) into M-matrices with, say, L nonsingular. For the remaining set of M-matrices, it is easy to see that not every singular and reducible $n \times n$ M-matrix admits an LU factorization into

* Received by the editors July 20, 1981.

† Institute for Computational Mathematics, Kent State University, Kent, Ohio 42242. The research of this author was supported in part by the Air Force Office of Scientific Research, and by the Department of Energy.

‡ Institute for Computational Mathematics, Kent State University, Kent, Ohio 44242. Visiting scholar from the Department of Applied Mathematics, Qing-Hua University, Beijing, People's Republic of China.

M-matrices with L nonsingular, as the particular matrix

$$A_1 = \begin{bmatrix} 0 & 0 \\ -1 & 1 \end{bmatrix} \tag{1.5}$$

directly shows. However, if $\mathcal{P}_n$ denotes the collection of all $n \times n$ permutation matrices, Kuo [5] has shown that, for any $n \times n$ M-matrix A, the subset of $\mathcal{P}_n$, defined by

$$\mathcal{P}_n^g(A) := \{P \in \mathcal{P}_n : PAP^T \text{ admits an } LU \text{ factorization into } M\text{-matrices with nonsingular } L\}, \tag{1.6}$$

is *never* empty. (Here, the superscript "g" in (1.6) refers to "good" permutations.) If $|\mathcal{P}_n^g(A)|$ denotes the cardinality of $\mathcal{P}_n^g(A)$ (i.e., the exact number of its elements), then the fact that $\mathcal{P}_n^g(A)$ is not empty implies (since $\mathcal{P}_n$ contains $n!$ elements) that

$$1 \leqq |\mathcal{P}_n^g(A)| \leqq n!, \tag{1.7}$$

for every $n \times n$ M-matrix A. From the above results, we remark that equality must evidently hold on the right in (1.7) for any nonsingular or irreducible $n \times n$ M-matrix A. From Funderlic and Plemmons [3], the same is true in (1.7) for any symmetric M-matrix A and for any M-matrix A for whicy $\underline{y}^T A \geqq 0$ for some $\underline{y} > 0$. Later (cf. (2.15)), we shall see that the first inequality in (1.7) is *sharp* for every $n \geqq 1$.

Our aim in this note is to determine upper and lower bounds for $|\mathcal{P}_n^g(A)|$, for any $n \times n$ M-matrix A. The outline of this note is as follows. We conclude this section with some needed notation, and in § 2, after giving some definitions, state our main results and give some applications of these results. The proofs of our main results are then given in § 3.

We assume, without loss of generality, that the $n \times n$ M-matrix A is in *normal reduced form* (cf. [7]), i.e.,

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,l} \\ & A_{2,2} & \cdots & A_{2,l} \\ & & \ddots & \vdots \\ 0 & & & A_{l,l} \end{bmatrix}, \tag{1.8}$$

where each diagonal submatrix $A_{j,j}$ is irreducible ($1 \leqq j \leqq l$). (As in [7], it is convenient to define all 1×1 null matrices here to be irreducible.) Of course, if A is irreducible, then $l = 1$ in (1.8). For large matrices, we remark that good software exists, for permuting the rows and columns of A to bring A into the form (1.8). For this, see George and Gustavson [4].

Next, if we define R_A as follows

$$R_A := \{j \text{ with } 1 \leqq j \leqq l : A_{j,j} \text{ is a singular and irreducible } M\text{-matrix}\}, \tag{1.9}$$

then $R_A \neq \varnothing$ if and only if A is a singular M-matrix. Continuing, we define the $l \times l$ upper triangular matrix $\mathcal{B}_A := [b_{i,j}]$, derived from A in (1.8) by means of

$$b_{i,j} := \begin{cases} 1 & \text{if } i \neq j \text{ and if } A_{i,j} \not\equiv \mathcal{O}, \\ 0 & \text{otherwise.} \end{cases} \tag{1.10}$$

Its directed graph $G_l(\mathcal{B}_A)$ on l vertices $V_1, V_2, \cdots, V_l$, is called the *block-directed graph* for the matrix A of (1.8). (As in [6] or [7], a *path* in $G_l(\mathcal{B}_A)$ from vertex V_i to vertex V_s is a sequence $\{b_{k_r,k_{r+1}}\}_{r=1}^{j}$ with $j \geqq 1$, $b_{k_r,k_{r+1}} \neq 0$, and with $k_1 = i$ and $k_{j+1} = s$.) For additional notation, with $\langle m \rangle := \{1, 2, \cdots, m\}$, let $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_k\}$ be a non-

empty subset of $\langle n\rangle$, and let $A[\alpha]$ denote the induced principal submatrix of the $n\times n$ matrix $A=[a_{i,j}]$, determined by α, i.e.,

$$A[\alpha]=[a_{i,j}],\quad \text{where } i,j\in\alpha. \tag{1.11}$$

As in [7], we shall say that α is a *proper* subset of $\langle n\rangle$ if $\varnothing\neq\alpha\subsetneq\langle n\rangle$.

If R_A of (1.9) is nonempty, then for each $j\in R_A$, we define the set

$$S_j := \{k\neq j\text{: there is a path in } G_1(\mathcal{B}_A) \text{ from vertex } V_k \text{ to vertex } V_j\}. \tag{1.12}$$

Because of the triangular form of (1.8), we note that S_j can contain only integers k satisfying $1\leqq k<j$, so that S_1, for example, is empty by definition. It is also convenient to say that

$$S_j \text{ is } \textit{full} \quad \text{iff} \quad S_j=\langle j-1\rangle. \tag{1.13}$$

2. Main results and applications. To state our first result, let A be an $n\times n$ singular M-matrix in normal reduced form (1.8) so that $R_A\neq\varnothing$, and suppose that S_j is not empty for some $j\in R_A$. Then, set

$$\mu := \max\{j\in R_A\text{: } S_j\neq\varnothing\}, \tag{2.1}$$

and *assume that S_μ is full.* Note, from (1.13), that the assumption that S_μ is full implies that $\mu>1$. With this value of μ, we define the following two principal submatrices of A, which are evidently M-matrices:

$$B := \begin{bmatrix} A_{1,1} & \cdots & A_{1,\mu} \\ & \ddots & \vdots \\ 0 & & A_{\mu,\mu} \end{bmatrix}, \qquad C := \begin{bmatrix} A_{1,1} & \cdots & A_{1,\mu-1} \\ & \ddots & \vdots \\ 0 & & A_{\mu-1,\mu-1} \end{bmatrix}. \tag{2.2}$$

This brings us to the statement of our first result, whose proof will be given in § 3.

THEOREM 1. (reduction algorithm). *Let A be an $n\times n$ M-matrix in normal reduced form* (1.8). *If $R_A=\varnothing$, or if $R_A\neq\varnothing$ and if $S_j=\varnothing$ for each $j\in R_A$* (cf. (1.12)), *then*

$$|\mathcal{P}_n^g(A)|=n!. \tag{2.3}$$

Otherwise, let μ be defined as in (2.1), *and assume that S_μ is full. If the matrix C of* (2.2) *is $s\times s$ and if $A_{\mu,\mu}$ is $m\times m$, then*

$$|\mathcal{P}_n^g(A)|=\frac{n!\cdot m|\mathcal{P}_s^g(C)|}{s!(s+m)}. \tag{2.4}$$

We remark that since the order of the matrix C of (2.2) is necessarily *less* than that of A, we can view Theorem 1 as a *reduction algorithm which* precisely relates $|\mathcal{P}_n^g(A)|$ for A to $|\mathcal{P}_s^g(C)|$ for the smaller matrix C. Of course, if C is nonsingular (so that its associated set R_C of (1.9) is empty), or if C is singular and its associated sets S_j of (1.12) are empty for all $j\in R_C$ (as is the case when C is irreducible), then $|\mathcal{P}_s^g(C)|=s!$, and the reduction algorithm necessarily *terminates.* Otherwise, the reduction algorithm can be continued if C satisfies the hypotheses of Theorem 1. Assuming that $R_A\neq\varnothing$, a sufficient condition that this reduction algorithm can be continued to termination is that

$$\text{for every } j\in R_A, \text{ either } S_j=\varnothing \text{ or } S_j=\langle j-1\rangle. \tag{2.5}$$

Now, if $R_A\neq\varnothing$, we further set

$$R_A^F := \{j\in R_A\text{: } S_j=\langle j-1\rangle\}, \tag{2.6}$$

and if $R_A^F \neq \varnothing$, we list (for convenience) its elements in decreasing order, i.e.,

$$R_A^F := \{\mu_1, \mu_2, \cdots, \mu_k\}, \quad \text{where } n \geqq \mu_1 > \mu_2 > \cdots > \mu_k \geqq 2. \tag{2.7}$$

For each $\mu_j \in R_A^F$, let t_j be defined as the order of the matrix B of (2.2) with $\mu = \mu_j$, and let s_j be similarly defined as the order of the matrix C of (2.2) with $\mu = \mu_j$. If m_j is the order of the matrix A_{μ_j,μ_j}, we note that $t_j = s_j + m_j$, $j = 1, 2, \cdots, k$.

With this notation, we state our next result, an extension of Theorem 1, whose proof will also be given in § 3.

THEOREM 2. *Let A be an $n \times n$ M-matrix, in normal reduced form* (1.8) *which satisfies* (2.5) *if $R_A \neq \varnothing$. If $R_A = \varnothing$, or if $R_A \neq \varnothing$ with $R_A^F = \varnothing$* (cf. (2.6)), *then*

$$|\mathscr{P}_n^g(A)| = n!. \tag{2.8}$$

Otherwise, with the above notation,

$$|\mathscr{P}_n^g(A)| = n! \frac{\prod_{j=1}^k m_j}{\prod_{j=1}^k t_j}, \tag{2.9}$$

where $R_A^F = \{\mu_1, \mu_2, \cdots, \mu_k\}$.

We next consider applications of Theorem 2. As our first application, consider the $n \times n$ singular reducible M-matrix (in normal reduced form)

$$D = \left[\begin{array}{c|c} D_{1,1} & D_{1,2} \\ \hline 0 & D_{2,2} \end{array}\right], \tag{2.10}$$

where $D_{1,1}$ is an $n_1 \times n_1$ nonsingular irreducible M-matrix, where $D_{2,2}$ is an $n_2 \times n_2$ singular irreducible M-matrix (with $n_1 + n_2 = n$), and where $D_{1,2} \neq \mathcal{O}$. In this example, $R_D = \{2\}$ and $S_2 = \{1\}$, so that the hypotheses of (2.5) are satisfied. Thus, $R_D^F = \{2\}$, so that $\mu_1 = 2$, $t_1 = n$, and $m_1 = n_2$. Applying (2.9) of Theorem 2 gives

$$|\mathscr{P}_n^g(D)| = n_2 \cdot (n-1)!, \tag{2.11}$$

the result of [7, Thm. 4]. Thus, Theorem 2 here generalizes [7, Thm. 4].

As a second application of Theorem 2, consider the following 12×12 singular reducible M-matrix (in normal reduced form)

$$E = \left[\begin{array}{cc|cc|c|cc|c|cc} 2 & -1 & 0 & -1 & & 0 & -1 & & 0 & -1 \\ -1 & 2 & -1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 \\ \hline 0 & & 1 & -1 & & 0 & -1 & & 0 & -1 \\ & & -1 & 1 & 0 & -1 & 0 & 0 & -1 & 0 \\ \hline 0 & & 0 & & 2 & -1 & 0 & & 0 & -1 \\ & & & & -1 & 2 & -1 & 0 & -1 & 0 \\ \hline 0 & & 0 & & 0 & & 1 & -1 & 0 & -1 \\ & & & & & & -1 & 1 & -1 & 0 \\ \hline 0 & & 0 & & 0 & & 0 & & 1 & -1 \\ & & & & & & & & -1 & 1 \\ \hline 0 & & 0 & & 0 & & 0 & & 2 & -1 \\ & & & & & & & & -1 & 2 \end{array}\right]. \tag{2.12}$$

In this case, (2.5) is satisfied, and $R_E = \{2, 4, 5\}$, and $R_E^F = \{2, 4\}$, so that $\mu_1 = 4$, $\mu_2 = 2$, $t_1 = 8$, $t_2 = 4$, and $m_2 = m_4 = 2$. Applying (2.9) yields

$$|\mathcal{P}_{12}^g(E)| = \frac{12!}{8}. \tag{2.13}$$

As our final application of Theorem 2, consider the particular upper triangular $n \times n$ singular M-matrix, defined by

$$H_n := \begin{bmatrix} 0 & -1 & -1 & \cdots & -1 & -1 \\ & 0 & -1 & \cdots & -1 & -1 \\ & & \ddots & \ddots & \vdots & \vdots \\ & & & & -1 & -1 \\ & & & & 0 & -1 \\ 0 & & & & & 0 \end{bmatrix}. \tag{2.14}$$

In this case, (2.5) is again satisfied, and $R_{H_n} = \{1, 2, \cdots, n\}$, $R_{H_n}^F = \{n, n-1, \cdots, 2\}$, and $m_j = 1$ for each $1 \leqq j \leqq n-1$. Applying (2.9) of Theorem 2 gives that

$$|\mathcal{P}_n^g(H_n)| = 1. \tag{2.15}$$

This example constructively shows that the first inequality in (1.7) is *sharp* for every $n \geqq 1$.

While Theorem 2 precisely determines $|\mathcal{P}_n^g(A)|$ for those $n \times n$ singular reducible M-matrices A satisfying (2.5), we next seek upper and lower bounds for $|\mathcal{P}_n^g(A)|$ for singular reducible M-matrices which do *not* satisfy (2.5).

Consider two $n \times n$ M-matrices A and B which are in normal reduced form

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,l} \\ & A_{2,2} & \cdots & A_{2,l} \\ & & \ddots & \vdots \\ 0 & & & A_{l,l} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_{1,1} & B_{1,2} & \cdots & B_{1,m} \\ & B_{2,2} & \cdots & B_{2,m} \\ & & \ddots & \vdots \\ 0 & & & B_{m,m} \end{bmatrix}, \tag{2.16}$$

where the diagonal submatrices $A_{j,j}$ and $B_{j,j}$ are irreducible. We say that A and B are *graph-compatible* if (cf. (2.16))

(2.17)
(i) $l = m$;
(ii) the order of $A_{j,j}$ is equal to the order of $B_{j,j}$ for each j with $1 \leqq j \leqq l$;
(iii) $R_A = R_B$ (cf. (1.9)).

With $S_j(A)$ and $S_j(B)$ denoting the sets of (1.12) associated with A and B when $R_A \neq \varnothing \neq R_B$, we come to:

PROPOSITION 3. *Let A and B be two $n \times n$ M-matrices which are graph-compatible* (cf. (2.17)). *If $R_A \neq \varnothing$ and if*

$$S_j(A) \subseteq S_j(B) \text{ for each } j \in R_A, \tag{2.18}$$

then

$$|\mathcal{P}_n^g(A)| \geqq |\mathcal{P}_n^g(B)|, \tag{2.19}$$

with strict inequality holding in (2.19) *if $S_j(A) \subsetneqq S_j(B)$ for some $j \in R_A$.*

We now use Proposition 3 as follows. Consider any $n \times n$ M-matrix A which is in normal reduced form (1.8). We shall construct two $n \times n$ singular reducible

M-matrices, $\underline{A}$ and $\bar{A}$, which are graph-compatible with A. Specifically, with

$$\underline{A} := \begin{bmatrix} \underline{A}_{1,1} & \underline{A}_{1,2} & \cdots & \underline{A}_{1,l} \\ & \underline{A}_{2,2} & \cdots & \underline{A}_{2,l} \\ & & \ddots & \vdots \\ 0 & & & \underline{A}_{l,l} \end{bmatrix}, \qquad \bar{A} := \begin{bmatrix} \bar{A}_{1,1} & \bar{A}_{1,2} & \cdots & \bar{A}_{1,l} \\ & \bar{A}_{2,2} & \cdots & \bar{A}_{2,l} \\ & & \ddots & \vdots \\ 0 & & & \bar{A}_{l,l} \end{bmatrix}, \tag{2.20}$$

we set

$$\begin{aligned} &\underline{A}_{i,i} = \bar{A}_{i,i} = A_{i,i} \quad \text{for each } 1 \leqq i \leqq l; \\ &\underline{A}_{i,j} = \bar{A}_{i,j} = A_{i,j} \quad \text{for each } j \notin R_A, \text{ and all } 1 \leqq i \leqq l; \\ &\underline{A}_{i,j} = \bar{A}_{i,j} = A_{i,j} \quad \text{for each } j \in R_A \text{ such that either } S_j(A) = \varnothing \\ &\qquad\qquad \text{or } S_j(A) = \langle j-1 \rangle, \text{ and all } 1 \leqq i \leqq l. \end{aligned} \tag{2.21}$$

Of course, if $R_A = \varnothing$ or if $R_A \neq \varnothing$ and A satisfies (2.5), then $\underline{A}$ and $\bar{A}$ are fully defined, with $\underline{A} = A = \bar{A}$. Otherwise, suppose there is a $j \in R_A$ for which $\varnothing \neq S_j(A) \subsetneqq \langle j-1 \rangle$. For such j's, we change zero blocks $A_{i,j}$ of A to nonzero blocks $\underline{A}_{i,j}$ in the upper triangular part of the jth column of A in such a way that

$$S_j(\underline{A}) = \langle j-1 \rangle \quad \text{for those } j \in R_A \text{ for which } \varnothing \neq S_j(A) \subsetneqq \langle j-1 \rangle. \tag{2.22}$$

Similarly, we change all nonzero blocks $A_{i,j}$, in the upper triangular part of the jth column of A, to be identically zero, thereby defining $\bar{A}_{i,j} \equiv \mathcal{O}$ for all $1 \leqq i < j$, so that

$$S_j(\bar{A}) = \varnothing \quad \text{for those } j \in R_A \text{ for which } \varnothing \neq S_j(A) \subsetneqq \langle j-1 \rangle. \tag{2.23}$$

Clearly, the matrices $\underline{A}$ and $\bar{A}$ are, by construction, *M*-matrices which are both graph-compatible with A. Moreover, if $R_A \neq \varnothing$, the matrices $\underline{A}$ and $\bar{A}$ are such that (2.5) is satisfied for each of these matrices, and also such that

$$S_j(\bar{A}) \subseteq S_j(A) \subseteq S_j(\underline{A}) \quad \text{for each } j \in R_A. \tag{2.24}$$

Now, $|\mathcal{P}_n^g(\bar{A})|$ and $|\mathcal{P}_n^g(\underline{A})|$ can be exactly computed from Theorem 2, so that from (2.19) of Proposition 3, we immediately have

THEOREM 4. *Let A be an $n \times n$ M-matrix in normal reduced form* (1.8). *With the $n \times n$ M-matrices $\underline{A}$ and $\bar{A}$ of* (2.21)–(2.23), *then either $R_A = \varnothing$ or $R_A \neq \varnothing$ and A satisfies* (2.5), *so that $\underline{A} = A = \bar{A}$ and*

$$|\mathcal{P}_n^g(\underline{A})| = |\mathcal{P}_n^g(A)| = |\mathcal{P}_n^g(\bar{A})|, \tag{2.25}$$

or $R_A \neq \varnothing$ and A does not satisfy (2.5), *so that*

$$|\mathcal{P}_n^g(\underline{A})| < |\mathcal{P}_n^g(A)| < |\mathcal{P}_n^g(\bar{A})|, \tag{2.26}$$

where $|\mathcal{P}_n^g(\underline{A})|$ and $|\mathcal{P}_n^g(\bar{A})|$ can be exactly determined from (2.9) *of Theorem* 2.

As an illustration of Theorem 4, consider the particular singular reducible *M*-matrix

$$J = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{2.27}$$

For this matrix, its associated graph-compatible matrices $\underline{J}$ and $\bar{J}$ can be taken to be

$$\underline{J} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \bar{J} = \begin{bmatrix} 1 & -1 & 0 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{2.28}$$

Here, the inequality of (2.26) of Theorem 4 can be computed to give

$$4 < |\mathcal{P}_4^g(J)| < 12. \tag{2.29}$$

By direct computation, we find, on the other hand, that $|\mathcal{P}_4^g(J)| = 8$.

As a final remark, suppose that an $n \times n$ M-matrix A is the direct sum of k M-matrices, i.e., in block-diagonal form,

$$A = \operatorname{diag}[A_{1,1}, A_{2,2}, \cdots, A_{k,k}], \tag{2.30}$$

where each $A_{j,j}$ is an $m_j \times m_j$ M-matrix. It is easy to see that

$$|\mathcal{P}_n^g(A)| = \frac{n! \prod_{i=1}^k |\mathcal{P}_{m_i}^g(A_{i,i})|}{\prod_{i=1}^k (m_i!)}, \quad \text{where } n = \sum_{i=1}^k m_i. \tag{2.31}$$

The point of this remark is that if (2.30) is valid, then Theorems 2 and 4 should be applied *only* to the matrices $A_{i,i}$, $1 \leqq i \leqq k$.

To illustrate this last remark, consider the following matrix

$$A = \left[\begin{array}{c|c|c|c} \square & \times & & \\ \hline & \Theta & & \\ \hline & & \Theta & \times \\ \hline & & & \Theta \end{array}\right] = \left[\begin{array}{c|c} A_{1,1} & \mathcal{O} \\ \hline \mathcal{O} & A_{2,2} \end{array}\right], \tag{2.32}$$

where $\square$, Θ denote respectively nonsingular and singular irreducible M-matrices, where blank blocks are identically zero, and where $\times$'s denote nonzero blocks. In this example, $R_A = \{2, 3, 4\}$, $S_2 = \{1\}$, $S_3 = \varnothing$, and $S_4 = \{3\}$; moreover, as $R_A \neq \varnothing$ and as (2.5) is not satisfied by A, Theorem 2 does not apply to A. However, A of (2.32) is the direct sum of the two matrices $A_{1,1}$ and $A_{2,2}$. As Theorem 2 *can* be applied to $A_{1,1}$ and $A_{2,2}$, then (2.31) can be applied, and $|\mathcal{P}_n^g(A)|$ can be precisely determined.

3. Proofs of results. For the convenience of the reader, we state below two results from [7] which will be used below.

THEOREM A. *Let A be an $n \times n$ M-matrix. Then, the following are equivalent:*

(i) *A admits an LU factorization into M-matrices with nonsingular L;*

(ii) *for every proper subset $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_k\}$ of $\langle n \rangle$ for which $A[\alpha]$ is singular and irreducible, there is no path in the directed graph $G_n(A)$ of A from vertex v_t to vertex v_{α_j} for any $t > \alpha_k$ and any $1 \leqq j \leqq k$.*

THEOREM B. *Let $A = [a_{i,j}]$ be an $n \times n$ M-matrix. Then, the following are equivalent:*

(i) *there exists an $x > 0$ such that $x^T A \geqq 0$;*

(ii) *$|\mathcal{P}_n^g(A)| = n!$;*

(iii) *for every proper subset $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_k\}$ of $\langle n \rangle$ for which $A[\alpha]$ is singular and irreducible, then $a_{t,p} = 0$ for all $t \notin \alpha$ and all $p \in \alpha$.*

Proof of Theorem 1. As (2.3) is immediate if $R_A = \varnothing$, assume first that $R_A \neq \varnothing$ and that the set S_j of (1.12) is empty for every $j \in R_A$. This implies that, for each $j \in R_A$, $A_{i,j} \equiv \mathcal{O}$ for each $1 \leqq i \leqq l$ with $i \neq j$. Since each singular and irreducible submatrix

of A of (1.8) must be some diagonal submatrix $A_{i,j}$ of A with $j \in R_A$, it follows that (iii) of Theorem B above is valid; whence, from (ii) of Theorem B, $|\mathcal{P}_n^g(A)| = n!$. This gives (2.3).

Next, assume that $S_j \neq \varnothing$ for some $j \in R_A$. With μ defined as in (1.13), we further assume that S_μ is full. The idea of the proof now is to use the equivalence of (i) and (ii) in Theorem A, in two stages, to deduce the desired result (2.4) of Theorem 1. We remark that if the directed graph $G_n(A)$ of the $n \times n$ M-matrix $A = [a_{i,j}]$ is associated with the n vertices $v_1, v_2, \cdots, v_n$, if σ is any permutation (1-1 transformation) on $\langle n \rangle$, and if its associated permutation matrix P_σ is defined by $P_\sigma = [\delta_{i,\sigma(j)}]$, then the directed graph $G_n(P_\sigma A P_\sigma^T)$ for $P_\sigma A P_\sigma^T$ is simply obtained by relabeling the vertices of $G_n(A)$ from v_j to $v_{\sigma(j)}$, while keeping all arcs intact. This observation will allow us to determine *which* rearrangements (permutations) of $\langle n \rangle$ are such that (ii) of Theorem A, applied to these rearrangements, is valid.

If the matrix B of (2.2) is $t \times t$, we first wish to establish that

$$|\mathcal{P}_n^g(A)| = |\mathcal{P}_t^g(B)| \cdot \frac{n!}{t!}. \tag{3.1}$$

Of course, if $t = n$, then $A = B$ and (3.1) trivially holds. Thus, we may assume that $t < n$. Consider any rearrangements of the first t positive integers, say $\{\nu_1, \nu_2, \cdots, \nu_t\}$, and consider any rearrangement $\{\tau_1, \tau_2, \cdots, \tau_{n-t}\}$ of the remaining positive integers $\{t+1, t+2, \cdots, n\}$. We then intersperse the integers of $\{\tau_1, \cdots, \tau_{n-t}\}$ among the integers of $\{\nu_1, \cdots, \nu_t\}$, thereby forming $\{\omega_1, \omega_2, \cdots, \omega_n\}$, a rearrangement of the first n integers, in such a way that $\{\omega_1, \cdots, \omega_n\} \backslash \{\tau_1, \cdots, \tau_{n-t}\} = \{\nu_1, \cdots, \nu_t\}$ and such that $\{\omega_1, \cdots, \omega_n\} \backslash \{\nu_1, \cdots, \nu_t\} = \{\tau_1, \cdots, \tau_{n-t}\}$. We claim that the number of ways of interspersing $\{\tau_1, \cdots, \tau_{n-t}\}$ with $\{\nu_1, \cdots, \nu_t\}$ is

$$K = \frac{n!}{t!}. \tag{3.2}$$

To see this, each distinct method of interspersing $\{\tau_1, \cdots, \tau_{n-t}\}$ among the integers of $\{\nu_1, \cdots, \nu_t\}$ applies equally well to each rearrangement $\{\nu_1', \cdots, \nu_t'\}$ of the first t positive integers. Thus, there are exactly the *same* number, say K, of ways of interspersing the integers of $\{\tau_1, \cdots, \tau_{n-t}\}$ among the integers of *each* rearrangement of the first t integers. Clearly, the totality of arrangements of $\{\omega_1, \cdots, \omega_n\}$ which can be obtained is, on one hand, $K \cdot t!$, while on the other hand, it is necessarily $n!$, which gives (3.2).

Next, we make the observation that if the rearrangement $\{\nu_1, \cdots, \nu_t\}$ corresponds to an element of $\mathcal{P}_t^b(B) := \langle n \rangle \backslash \mathcal{P}_t^g(B)$, then it is easily seen that every interspersing of the integers of any rearrangement $\{\tau_1, \cdots, \tau_{n-t}\}$ by definition corresponds to an element of $\mathcal{P}_n^b(A)$. Thus, to obtain a rearrangement $\{\omega_1, \cdots, \omega_n\}$ in $\mathcal{P}_n^g(A)$, it is necessary to *begin* with a rearrangement $\{\nu_1, \cdots, \nu_t\}$ which is in $\mathcal{P}_t^g(B)$, followed by *any* interspersing of *any* $\{\tau_1, \cdots, \tau_{n-t}\}$. (The reason that this is valid is that the $n-t$ integers $\{\tau_1, \cdots, \tau_{n-t}\}$ necessarily correspond to vertices in the directed graph $G_n(A)$ of A which, by construction, have *no* path to the singular irreducible submatrix $A_{\mu,\mu}$, and hence play no role in applying (ii) of Theorem A to $A_{\mu,\mu}$.) Thus, using (3.2), $|\mathcal{P}_n^g(A)|$ is given by

$$|\mathcal{P}_n^g(A)| = \frac{n!}{t!} |\mathcal{P}_t^g(B)|. \tag{3.3}$$

We now relate, in the second part of the proof, the quantities $|\mathcal{P}_t^g(B)|$ and $|\mathcal{P}_s^g(C)|$, where B and C are defined in (2.2). Since S_μ is full by hypothesis, then $S_\mu = \langle \mu - 1 \rangle$.

By definition, the matrices $A_{\mu,\mu}$, C and B (cf. (2.2)) are respectively of orders m, s and t, with $t = m + s$. In analogy to the first part of the proof, we consider any rearrangement $\{\tau_1, \cdots, \tau_s\}$ of the first s positive integers, and any rearrangement $\{\eta_1, \cdots, \eta_m\}$ of the integers $\{s+1, \cdots, t\}$, and we intersperse $\{\eta_1, \cdots, \eta_m\}$ among the integers of $\{\tau_1, \cdots, \tau_s\}$, thereby forming $\{\omega_1, \omega_2, \cdots, \omega_t\}$. As before, to obtain an element in $\mathscr{P}_t^g(B)$, it is necessary to *begin* with a rearrangement $\{\tau_1, \cdots, \tau_s\}$ which is in $\mathscr{P}_s^g(C)$. Moreover, because by hypothesis $S_\mu = \langle \mu - 1 \rangle$, we see from (ii) of Theorem A that the final element ω_t of $\{\omega_1, \cdots, \omega_t\}$ in $\mathscr{P}_t^g(B)$ *must* be from $\{\eta_1, \cdots, \eta_m\}$. For each fixed $\{\tau_1, \cdots, \tau_s\}$ in $\mathscr{P}_s^g(C)$, it is easily seen that there are the *same* number, namely $m \cdot (t-1)!/s!$ of such interspersings of $\{\eta_1, \cdots, \eta_m\}$, such that the last element ω_t is from $\{\eta_1, \cdots, \eta_m\}$. Thus,

$$|\mathscr{P}_t^g(B)| = \frac{|\mathscr{P}_s^g(C)| \cdot m \cdot (t-1)!}{s!}. \tag{3.4}$$

If we combine (3.4) with (3.3), we obtain (since $t = s + m$) the desired result (2.4). □

Proof of Theorem 2. As (2.8) is immediate if $R_A = \varnothing$, assume first that $R_A \neq \varnothing$ and that $R_A^F = \varnothing$. But, $R_A^F = \varnothing$ implies from (2.5) that $S_j = \varnothing$ for each $j \in R_A$, which with (2.3) of Theorem 1 gives that $|\mathscr{P}_n^g(A)| = n!$ in (2.8). Hence, we may assume that R_A^F is not empty, so that from (2.7), $R_A^F = \{\mu_1, \mu_2, \cdots, \mu_k\}$ where $n \geqq \mu_1 > \mu_2 > \cdots > \mu_k \geqq 2$. Now, let $C^{(j)}$ denote the matrix C of (2.2) when $\mu = \mu_j$, $j = 1, 2, \cdots, k$. In addition, we set

$$C^{(0)} := A. \tag{3.5}$$

From the discussion preceding Theorem 2, s_j denotes the order of each $C^{(j)}$, so that $s_0 := n$. Similarly, m_j denotes the order of the matrix A_{μ_j,μ_j}. Then, applying (2.4) of Theorem 1 to $C^{(j)}$ yields

$$|\mathscr{P}_{s_j}^g(C^{(j)})| = \frac{(s_j)! m_{j+1} |\mathscr{P}_{s_{j+1}}^g(C^{(j+1)})|}{(s_{j+1})!(t_{j+1})}, \qquad j = 0, 1, \cdots, k-1, \tag{3.6}$$

where $t_j := s_j + m_j$. On multiplying the quantities of (3.6) for all $j = 0, 1, \cdots, k-1$, we obtain (since $s_0 := n$)

$$|\mathscr{P}_n^g(A)| = \frac{n! \prod_{j=1}^k m_j |\mathscr{P}_{s_k}^g(C^{(k)})|}{(s_k)! \cdot \prod_{j=1}^k t_j}. \tag{3.7}$$

But, since the irreducible diagonal submatrices of $C^{(k)}$ are either nonsingular, or singular with associated sets S_j empty, from (2.5) and (2.7), $|\mathscr{P}_{s_k}^g(C^{(k)})| = (s_k)!$, and (3.7) then reduces to the desired result (2.9). □

Proof of Proposition 3. With the hypotheses of Proposition 3, consider any permutation P in $\mathscr{P}_n^g(B)$. From the equivalence of (i) and (ii) in Theorem A, it is easy to verify that the hypothesis of graph-compatibility and the inclusions of (2.18) imply that P is also in $\mathscr{P}_n^g(A)$, whence $|\mathscr{P}_n^g(A)| \geqq |\mathscr{P}_n^g(B)|$, the desired inequality of (2.19).

Next, suppose that there is a $j \in R_A$ for which $S_j(A) \subsetneq S_j(B)$, along with $S_k(A) \subseteq S_k(B)$ for each $k \in R_A$, and let

$$s := \max\{k : k \in S_j(B) \backslash S_j(A)\}, \quad \text{where } s < j. \tag{3.8}$$

By definition, there is a path from vertex V_s to vertex V_j in the block-directed graph for the matrix B, but no such path in the block-directed graph for the matrix A. With A in reduced normal form (2.16) and with $\alpha := \{s, s+1, \cdots, j\}$, set

$$T_s^A := \{t : s \leqq t \leqq j \text{ and there is a path from } V_s \text{ to } V_j \text{ for } A\} \cup \{s\}. \tag{3.9}$$

By definition,

$$s \in T_s^A \quad \text{and} \quad j \in \{\alpha \backslash T_s^A\}. \tag{3.10}$$

Then, as in the proof of Theorem 1, with each vertex V_i, we associate n_i positive integers (where $A_{i,i}$ and $B_{i,i}$ are $n_i \times n_i$ integers), numbered consecutively so that the integers associated with V_1 are $\{1, 2, \cdots, n_1\}$, those associated with V_2 are $\{n_1+1, \cdots, n_1+n_2\}$, etc. Now, alter $\langle n \rangle$ (thereby forming a rearrangement of $\langle n \rangle$) by simply removing the consecutive integers, corresponding in sequence to the vertices V_l in T_s^A, and placing them (without changing their relative positions) immediately after the last integer associated with the last vertex of V_j. Using (3.10), Theorem A, and the fact that there *is* a path from vertex V_s to vertex V_j for the matrix B, this new rearrangement of $\langle n \rangle$ can be seen to be in the set $\mathscr{P}_n^g(A)$, but not in the set $\mathscr{P}_n^g(B)$. Thus, $|\mathscr{P}_n^g(A)| > |\mathscr{P}_n^g(B)|$, which gives strict inequality in (2.19). □

Proof of Theorem 4. The proof of Theorem 4 follows easily from Proposition 3. First, if $R_A = \varnothing$ or if $R_A = \varnothing$ and A satisfies (2.5), the construction of $\underline{A}$ and $\bar{A}$ is such that $\underline{A} = A = \bar{A}$ in this case, from which (2.25) follows. Otherwise, assume $R_A \neq \varnothing$ and that A does not satisfy (2.5). Hence, there exists a $j \in R_A$ for which $\varnothing \neq S_j(A) \subsetneqq \langle j-1 \rangle$. For this j, the construction of $\underline{A}$ and $\bar{A}$ from (2.22)–(2.23) shows that

$$S_j(\bar{A}) = \varnothing \neq S_j(A) \subsetneqq S_j(\underline{A}) = \langle j-1 \rangle, \tag{3.11}$$

as well as

$$S_k(\bar{A}) \subseteq S_k(A) \subseteq S_k(\underline{A}) \quad \text{for which } k \in R_A. \tag{3.12}$$

Thus, strict inequality holds in (2.19) of Proposition 3, i.e.,

$$|\mathscr{P}_n^g(\underline{A})| < |\mathscr{P}_n^g(A)| < |\mathscr{P}_n^g(\bar{A})|, \tag{3.13}$$

which gives the desired result (2.26) of Theorem 4. □

REFERENCES

[1] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[2] M. FIEDLER AND V. PTÁK, *On matrices with nonpositive off-diagonal elements and positive principal minors*, Czech. Math. J., 12 (1962), pp. 382–400.

[3] R. E. FUNDERLIC AND R. J. PLEMMONS, *LU decompositions of M-matrices by elimination without pivoting*, Linear Algebra Appl., to appear.

[4] A. GEORGE AND F. G. GUSTAVSON, *A new proof on permuting to block triangular form*, Ibid., to appear.

[5] I.-WEN KUO, *A note on factorization of singular M-matrices*, Ibid., 16 (1977), pp. 217–220.

[6] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[7] R. S. VARGA AND D.-Y. CAI, *On the LU factorization of M-matrices*, Numer. Math., 38 (1981), pp. 179–192.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 2, June 1982

0196-5212/82/0302-0012 $01.00/0

ISOMORPHISM TESTING IN HOOKUP CLASSES*

M. M. KLAWE†, D. G. CORNEIL‡ AND A. PROSKUROWSKI§

Abstract. Hookup classes are classes of graphs with a certain type of recursive definition, which can be viewed as a generalization of k-trees. We show that many hookup classes of graphs are isomorphism complete, and give polynomial isomorphism algorithms for the others. Other results in this paper include the development of a structural decomposition for hookup graphs and similar isomorphism results for generalizations of hookup classes, including a polynomial isomorphism testing algorithm for chordal graphs with bounded maximum clique size.

1. Introduction. The objective of this paper is to study the computational complexity of isomorphism testing in classes of "hookup" graphs, which are classes of graphs with a particular type of recursive definition. For two graphs A and G, we define the *hookup class* of A and G, denoted by $[A, G]$, as follows. A graph H belongs to $[A, G]$ if either H is isomorphic to A or there exists a vertex z of H such that the subgraph of H induced by the neighborhood of z is isomorphic to G and the graph $H\backslash\{z\}$ belongs to $[A, G]$. Another way of describing the graphs in $[A, G]$ is that they are the graphs that can be obtained by starting with a copy of A and adjoining vertices one by one in such a way that whenever a vertex is adjoined it is made adjacent to every vertex of (or "hooked up" to) a copy of G in the preceding graph. We call A the initial graph, and G the hooking graph, of the hookup class $[A, G]$. Some well-known examples of hookup classes are $[K_k, K_k]$ for $k \geqq 1$, since this is simply the class of k-trees (see [14], [15], [17]); in particular, for $k=1$ this hookup class is simply the family of trees.

The general problem of determining whether two graphs are isomorphic has become one of the most tantalizing open problems in the field of computational complexity. Despite considerable effort, the problem has neither been shown to be polynomial nor shown to be NP-complete. The many practical applications of graph isomorphism merely add to the interest of this problem.

Recent results in this area can be split into two groups: those which provide a polynomial or subexponential isomorphism testing algorithm for some particular class of graphs and those which show that a particular class of graphs is isomorphism complete, i.e., if there is a polynomial algorithm for isomorphism testing in this particular class, then there is a polynomial algorithm for isomorphism testing of all graphs. For example, polynomial algorithms for isomorphism testing are known for trees and, more generally, k-trees for fixed k [8]; planar graphs and, more generally, graphs of bounded genus [5], [9], [10], [12]; and, most recently, graphs of bounded valence [11]. On the other hand, classes of graphs which are known to be isomorphism complete include regular graphs, chordal graphs, minimally 2-connected graphs, line graphs and self-complementary graphs (see [1] for a recent review of isomorphism complete problems).

The major result of this paper is to show that for any graphs A and G, the hookup class $[A, G]$ is either isomorphism complete or has a polynomial isomorphism testing

* Received by the editors February 24, 1981, and in revised form July 31, 1981. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada under grant A7671.

† IBM Research, 5600 Cottle Rd., San Jose, California 95193.
‡ University of Toronto, Toronto, Ontario, Canada M5S 1A7.
§ University of Oregon, Eugene, Oregon, 97403.

algorithm. We also show that for any graph G which is not a complete graph there is a graph A with $|A| \leqq 3|G|$ such that $[A, G]$ is isomorphism complete. (Here, $|A|$ denotes the number of vertices in A.) Conversely, if G is a complete graph, then isomorphism testing in $[A, G]$ is polynomial for every graph A. In addition, we are able to give a complete characterization of the graphs G such that $[G, G]$ is isomorphism complete. These results depend on obtaining a structural decomposition of hookup graphs which is (almost) invariant with respect to isomorphism.

In § 2 the previously mentioned decomposition of hookup graphs is presented along with a polynomial algorithm for obtaining it. The results showing that certain hookup classes are isomorphism complete are given in § 3, as well as the characterization of graphs G such that $[G, G]$ is isomorphism complete. Section 4 develops the polynomial isomorphism algorithm for the remaining hookup classes, and concludes by showing that the class of all k-trees is isomorphism complete which contrasts the polynomial isomorphism algorithm for k-trees when k is fixed. The last section considers generalizations of hookup classes. In particular we show that if vertices are allowed to hook up to subgraphs which are only partially isomorphic to G then the hookup class obtained is isomorphism complete whenever G is not a complete graph and the class is not trivial. The paper concludes with a sketch of a polynomial isomorphism algorithm for var-k-trees, which are a generalization of k-trees in which vertices are allowed to hook up to any complete subgraph of size $\leqq k$. An alternate characterization of var-k-trees is as the class of connected chordal graphs with maximum clique size $\leqq k+1$.

2. Preliminaries. This section is devoted to developing a structural decomposition for hookup graphs, which will be used in obtaining polynomial algorithms for some hookup classes. After defining this decomposition and proving some facts about it, we give an algorithm for obtaining it which is linear in the size of the hookup graph.

For convenience, here and elsewhere in this paper we will confuse a graph with its set of vertices. Thus H may refer either to the graph itself or merely to $V(H)$. We will denote the neighborhood of a vertex x in H by Γ_x or by $\Gamma_x(H)$ when we wish to specify which graph we mean. Thus $\Gamma_x(H)$ denotes the set of vertices of H which are adjacent to x in H. For two graphs F and H, we use $F \simeq H$ to denote that F is isomorphic to H.

We recursively define a *base* of a hookup graph as follows. If H belongs to $[A, G]$ then a subgraph B of H is a base of H if B is isomorphic to A and either $H = B$ or there exists a vertex z of H such that $\Gamma_z \simeq G$, $H\backslash\{z\} \in [A, G]$ and B is a base of $H\backslash\{z\}$. Intuitively, B is a base of H if B could have been used as the initial graph in a vertex by vertex construction of H. It is easy to see that a hookup graph can have many different bases but that every hookup graph must have at least one base. We let $[A, G]_b$ denote the set of pairs (H, B) such that $H \in [A, G]$ and B is a base of H. If (H, B) and (H', B') belong to $[A, G]_b$, then we say they are isomorphic if there exists an isomorphism $\psi: H \to H'$ such that $\psi(B) = B'$, and refer to ψ as a base-preserving isomorphism.

For (H, B) in $[A, G]_b$, we define the B-decomposition of H to be the sequence of sets $B(0), B(1), \cdots, B(p)$, where $B(0) = B$, and $B(i+1)$ is recursively defined by

$$B(i+1) = \left\{ x \in H \Big/ \bigcup_{j=0}^{i} B(j) : \left(\Gamma_x \cap \bigcup_{j=0}^{i} B(j) \right) \simeq G \right\}$$

and $p = \max\{k : B(k) \neq \varnothing\}$.

If $x \in B(k)$ for $k \geqq 1$, we denote $\Gamma_x \cap \bigcup_{j=0}^{k-1} B(j)$ by $G(x)$ and call this the *support* of x in (H, B). Intuitively, the B-decomposition indicates the order in which vertices could be "hooked-up" in a vertex by vertex construction of H with B as the initial graph. All the vertices in $B(i)$ must be hooked-up before any vertex in $B(i+1)$, and within B(i) the vertices may be hooked-up in arbitrary order. In this interpretation $G(x)$ is the copy of G to which x hooks up.

The following lemma provides the basic facts which we will need about a B-decomposition.

LEMMA 2.1. *Let (H, B) belong to $[A, G]_b$. Then the B-decomposition $B(0)$, $B(1), \cdots B(p)$, of H has the following properties*:

(a) *If $i \neq j$, then $B(i) \cap B(j) = \varnothing$.*

(b) *If $x \in B(i)$ for $i \geqq 1$, then $G(x) \simeq G$, i.e., $\Gamma_x \cap \bigcup_{j=0}^{i-1} B(j) \simeq G$.*

(c) *If $x \in B(i)$ for $i \geqq 2$, then $\Gamma_x \cap \bigcup_{j=0}^{i-2} B(j) \not\simeq G$.*

(d) $H = \bigcup_{j=0}^{p} B(j)$.

(e) *If $x, y \in B(i)$ for $i \geqq 1$ and $x \neq y$, then x and y are not adjacent.*

(f) *If $x \in B(i)$ for $i \geqq 1$, then $\Gamma_x \cap \bigcup_{j=0}^{i} B(j) \simeq G$.*

Proof. Properties (a), (b) and (c) follow directly from the definition of B-decomposition, and (f) follows immediately from (b) and (e). Properties (d) and (e) are proved by induction on $|H|$. If $H = B$, then (d) and (e) are trivially satisfied, so we may assume that H has a vertex z such that $\Gamma_x \simeq G$ and B is a base of $H\backslash\{z\}$, and that the lemma is true for the B-decomposition of $H\backslash\{z\}$, which we denote by $C(0), C(1), \cdots, C(q)$. First notice that if z is not in $\bigcup_{j=0}^{p} B(j)$ then we must have $B(j) = C(j)$ for each j, and $p = q$. But then, since $H\backslash\{z\} = \bigcup_{j=0}^{q} C(j)$ by the inductive assumption, we have that $\Gamma_z \subset H\backslash\{z\} = \bigcup_{j=0}^{p} B(j)$. Hence z must be in $B(t)$ where $t = \min\{i : \Gamma_z \subset \bigcup_{j=0}^{i-1} B(j)\} \leqq p$, a contradiction. Thus we may assume that z belongs to $B(t)$ for some $t \geqq 1$. Now, since $\Gamma_z \simeq G \simeq \Gamma_z \cap \bigcup_{j=0}^{t-1} B(j)$, it is clear that $\Gamma_z \cap \bigcup_{j=t}^{p} B(j) = \varnothing$, which shows that z does not belong to Γ_x for any x in $\bigcup_{j=t}^{p} B(j)$. Thus, for $j \neq t$ we have $B(j) = C(j)$ and $B(t) = C(t) \cup \{z\}$. As $H\backslash\{z\} = \bigcup_{j=0}^{q} C(j)$, obviously $H = \bigcup_{j=0}^{p} B(j)$. Moreover, for distinct x, y in $B(i)$ for $i \geqq 1$, we either have both x and y in $C(i)$, and hence x is not adjacent to y by the inductive assumption or $i = t$ and one of x or y is z, in which case x and y cannot be adjacent because, as noted above, $\Gamma_x \cap B(t) = \varnothing$. □

For (H, B) in $[A, G]_b$, we will find it useful to define the *level function* with respect to B, $l(x)$, on the vertices of H, by $l(x) = i$ if $x \in B(i)$. We now present an algorithm which, taking a graph H and subgraph B as inputs, determines whether (H, B) belongs to $[A, G]_b$ and, if so, obtains the B-decomposition and the sets $G(x)$ for each x in $H\backslash B$. By using appropriate data structures, the algorithm can be implemented so that its running time is linear in $|H|$, though the constant factor depends factorially on $|A|$ and $|G|$.

ALGORITHM 2.2.

INPUT: A graph H and induced subgraph B.

The value of the logical variable FAIL is set to TRUE during the execution of the algorithm as soon as it is determined that (H, B) is not in $[A, G]_b$. During the execution of the algorithm, for each $x \in H\backslash B$ its set of adjacent vertices is always partitioned into two sets, B-ADJ (x) and its complement REST-ADJ (x).

The sets $B(0), B(1), \cdots, B(p)$ will be obtained iteratively beginning with $B(0)$, which is just B itself. At the time that $B(i)$ is being formed, if a vertex x is not in any $B(j)$ for $j < i$, then B-ADJ (x) contains $\Gamma_x \cap \bigcup_{j=0}^{i-1} B(j)$. If $x \in B(j)$ for some j with $1 \leqq j < i$, then B-ADJ (x) contains $G(x)$. Thus, at the end of execution if FAIL = FALSE, the sets B-ADJ (x) will be exactly the desired sets $G(x)$.

IF B is isomorphic to A THEN FAIL := FALSE;
ELSE FAIL := TRUE;
IF FAIL = FALSE THEN DO;
Initialize $B(0)$ to be B and all other $B(i)$ to be empty;
For each $x \in H \backslash B$ DO;
initialize B-ADJ $(x) = \Gamma_x \cap B$;
initialize REST-ADJ $(x) = \Gamma_x \backslash$B-ADJ (x);
END;
$I := 0$;
END;
DO UNTIL ($B(I)$ is empty or FAIL = TRUE);
$I := I + 1$;
For each $x \in H \backslash \cup_{j=0}^{I-1} B(j)$ with $|\text{B-ADJ}(x)| = |G|$
IF B-ADJ $(x) \simeq G$ THEN add x to $B(I)$;
ELSE FAIL := TRUE;
For each $x \in B(I)$
For each $y \in$ REST-ADJ (x)
IF $y \in B(I)$ THEN FAIL := TRUE;
ELSE move x from REST-ADJ (y) to B-ADJ (y);
END;
$p := I - 1$;
IF $H \backslash \cup_{j=0}^{p} B(j)$ is not empty THEN FAIL := TRUE;
SUCCESS := $\neg$FAIL;
OUTPUT(SUCCESS);

To see that the algorithm performs correctly, it is easy to see from Lemma 2.1 that if (H, B) is in $[A, G]_b$, the algorithm will indeed output TRUE. On the other hand, it is straightforward to give a proof by induction on $|H|$ that if the algorithm outputs TRUE then (H, B) does belong to $[A, G]_b$.

To be efficient, data structures for REST-ADJ (y) and B-ADJ (y) must allow x to be moved from REST-ADJ (y) to B-ADJ (y) in constant time, which is easily accomplished by the use of double pointers between the record for x in REST-ADJ (y) and that for y in REST-ADJ (x). Also, to avoid searching $H \backslash \cup_{j=0}^{I-1} B(j)$ each time to find those vertices x such that $|\text{B-ADJ}(x)| = |G|$, this set can be maintained by the use of variables DEG-B (x) to count $|\text{B-ADJ}(x)|$ and checking whether DEG-B $(x) = |G|$ every time that DEG-B (x) is incremented. Finally, by adding variables $L(x)$ which record the level $l(x)$ as soon as it is determined, it is possible to check whether $y \in B(I)$ in constant time. Using this type of implementation, it is easy to check that the algorithm is linear in $|E(H)|$, the number of edges of H, and hence in $|H|$, since clearly $|E(H)| \leqq |A|(|A|-1) + |G|(|H|-|A|)$.

3. Isomorphism completeness of hookup classes. In this section we introduce the notion of G-bolt which will be used to show that many hookup classes are isomorphism complete. Let G and H be graphs. We say that H contains a G-bolt if there exist vertices x and y of H and induced subgraphs X, Y and Z of H such that

(3.1.1) x and y are not adjacent,
(3.1.2) $x, y \in Z$,
(3.1.3) $X, Y, Z \simeq G$,
(3.1.4) $X \cap (Z \backslash \{x\}) = \Gamma_x \cap Z$ and $Y \cap (Z \backslash \{y\}) = \Gamma_y \cap Z$.

We denote the *G-bolt* by the 5-tuple (x, y, X, Y, Z).

THEOREM 3.2. *If there exists an $H \in [A, G]$ containing a G-bolt, then the class $[A, G]$ is isomorphism complete.*

Proof. We will show that for any graph F we can construct in polynomial time a graph $h(F)$ such that $h(F) \in [A, G]$, $|h(F)| = O(|F|^2)$, and moreover, for any graphs F and F', the graphs $h(F)$ and $h(F')$ are isomorphic if and only if F and F' are.

Let $H \in [A, G]$ contain a G-bolt (x, y, X, Y, Z), and let $n = |H|$ and $m = |F|$. For convenience we assume $m \geqq 2$. The graph $h(F)$ is constructed as follows. First form a graph $f(F)$ by adjoining to H two sets of vertices $\{x(v): v \in F\}$ and $\{y(v): v \in F\}$ such that $\Gamma_{x(v)} = X$ and $\Gamma_{y(v)} = Y$ for all $v \in F$. Since $X \simeq G$ and $Y \simeq G$, it is easy to see that $f(F) \in [A, G]$. Now adjoin the set of vertices $\{w(v, i): v \in F, 1 \leqq i \leqq 2m + n\}$ to $f(F)$ to form a graph $g(F)$, setting $\Gamma_{w(v,i)} = (Z \backslash \{x, y\}) \cup \{x(v), y(v)\}$ for $v \in F$, $1 \leqq i \leqq 2m + n$. Note that $\Gamma_{w(v,i)} \simeq Z$ for each v and i, since $\Gamma_{x(v)} \cap (Z \backslash \{x\}) = X \cap (Z \backslash \{x\}) = \Gamma_x \cap Z$ and $\Gamma_{y(v)} \cap (Z \backslash \{y\}) = Y \cap (Z \backslash \{y\}) = \Gamma_y \cap Z$ by property (3.1.4) of G-bolts. Now as $Z \simeq G$ we have $\Gamma_{w(v,i)} \simeq G$, and hence it is easy to see that $g(F) \in [A, G]$. Finally, form $h(F)$ by adjoining the set of vertices $\{z(v, u, i): u$ is adjacent to v in F, $1 \leqq i \leqq n\}$ to $g(F)$, so that $\Gamma_{z(v,u,i)} = Z \backslash \{x, y\}) \cup \{x(v), y(u)\}$ for $1 \leqq i \leqq n$. Notice that when u and v are adjacent in F both $z(v, u, i)$ and $z(u, v, i)$ are adjoined for each i. As before $\Gamma_{z(v,u,i)} \simeq G$ for all v, u, i, and, hence $h(F) \in [A, G]$. Moreover $|h(F)| = n + 2m + m(2m + n) + 2ne$, where e is the number of edges in F. Thus, clearly $|h(F)| = O(|F|^2)$.

To see that this construction preserves isomorphism notice that F can be reconstructed from $h(F)$ as follows. The set of vertices $\{x(v), y(v): v \in F\}$ can be identified by their degree, since letting d be the maximum degree of f we have

(i) for $z \in H \backslash (X \cup Y \cup (Z \backslash \{x, y\}))$, degree $(z) < n$;

(ii) for $z \in X \cup Y \backslash (Z \backslash \{x, y\})$, degree $(z) < 2m + n$;

(iii) for $z = z(v, u, i)$ or $z = w(v, i)$, degree $(z) = |G| < n$;

(iv) for $z \in (Z \backslash \{x, y\})$, since $m \geqq 2$ and $e \geqq d$

$$\text{degree}(z) \geqq m(2m + n) + 2en \geqq 4m + 2n + 2dn;$$

(v) for $z = x(v)$ or $y(v)$

$$2m + n \leqq \text{degree } (z) \leqq 2m + n + |G| + dn \leqq 2m + 2n + dn.$$

For r and s, two vertices of $h(F)$, let $c(r, s)$ be the number of vertices of $h(F)$ which are adjacent to both r and s. Then it is easy to check that we have, for any distinct u and v in F,

(vi) $c(x(u), x(v)) = c(y(u), y(v)) = |G| < n$;

(vii) $c(x(u), y(v)) = |X \cap Y|$, if u and v are not adjacent in F;

(viii) $c(x(u), y(v)) = n + |X \cap Y|$ if u and v are adjacent in F;

(xi) $c(x(v), y(v)) = 2m + n + |X \cap Y|$.

Thus it is clear that F may be reconstructed by examining the function c on the set of vertices $\{x(v), y(v): v \in F\}$. Therefore $h(F) \simeq h(F')$ if and only if $F \simeq F'$, and hence, a polynomial isomorphism testing algorithm for graphs in $[A, G]$ would yield a polynomial isomorphism testing algorithm for all graphs. □

The fact that many hookup classes are isomorphism complete follows from the next theorem.

THEOREM 3.3. *If G is not a complete graph, then there exists a graph A such that $[A, G]$ is isomorphism complete.*

Proof. By Theorem 3.2 it suffices to construct a graph A which contains a G-bolt. Since G is not complete, there exists a pair of nonadjacent vertices x, y in G. Let G_1, G_2 and G_3 be three disjoint copies of G, and let x, x_2, x_3 and y_1, y_2, y_3 be the vertices corresponding to x and y in these copies of G. Let A be the graph formed by first

identifying x_1 and its neighborhood in G_1 with x_2 and its neighborhood G_2 and then identifying y_1 and its neighborhood in G_1 with y_3 and its neighborhood in G_3. If we take Z, X and Y to be the subgraphs of A induced by the vertices from G_1, G_2 and G_3 respectively, then it is easy to see that (x_1, y_1, X, Y, Z) form a G-bolt in A. □

Figure 1 shows A when G is the 5-cycle.

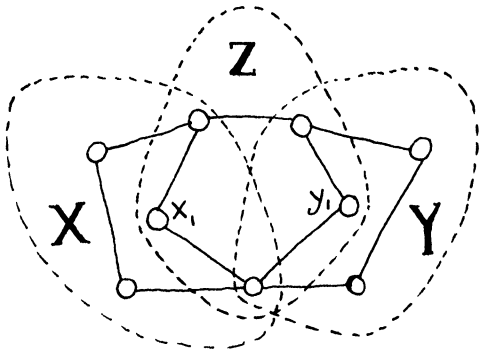


FIG. 1

We now attack the problem of characterizing the graphs G such that $[G, G]$ has a graph containing a G-bolt. We begin with some notation and an easy lemma.

For graphs H and F, let $H \equiv F$ denote the graph consisting of the union of the graphs H and F together with all possible edges joining vertices of H to vertices of F. Note that F may be the empty graph, in which case, $H \equiv F$ is simply H.

LEMMA 3.4. *Let G be a graph such that $[G, G]$ has a graph containing a G-bolt, and let F be any graph. Then the class $[G \equiv F, G \equiv F]$ has a graph containing a $(G \equiv F)$-bolt.*

Proof. It can easily be shown by induction on $|H|$ that whenever $H \in [G, G]$ we have $H \equiv F \in [G \equiv F, G \equiv F]$. Moreover, if H contains a G-bolt (x, y, X, Y, Z) then $(x, y, X \equiv F, Y \equiv F, Z \equiv F)$ is a $(G \equiv F)$-bolt in $H \equiv F$. □

The family of chordal graphs can be defined recursively by saying that a graph C is chordal if either C is complete or there exists $z \in C$ such that $\Gamma_z \simeq K_p$ for $p \geqq 0$ and $C \backslash \{z\}$ is chordal. Note that by allowing $p = 0$ a chordal graph is not necessarily connected. It is well known that a graph C is chordal if and only if each cycle of length at least four has a chord, and in fact this is often used as the definition. Let $\alpha(C)$ denote the size of the largest clique in C. We should note here that we use the term clique to refer to a complete subgraph which is maximal with respect to that property. We say that graph G is *chordal-clique-complete* if $G = C \equiv K_{\alpha(C)-1}$, where C is a noncomplete chordal graph. Finally, a graph G is *chordal-clique-complete-extended* if $G = G' \equiv F$, where G' is chordal-clique-complete and F is any graph. The following theorem completely characterizes those graphs G such that $[G, G]$ has a graph containing a G-bolt.

THEOREM 3.5. *The class $[G, G]$ has a graph containing a G-bolt if and only if G is chordal-clique-complete-extended.*

The rest of this section will be devoted to proving this theorem. We begin with some definitions and elementary facts which will be used to show that if G is chordal-clique-complete, then $[G, G]$ has a graph containing a G-bolt.

A *proper k-tree* is any graph belonging to $[K_k, K_k]$ other than K_k itself. Obviously, proper k-trees are chordal graphs. A *simplicial point* of a chordal graph is a vertex x

such that $\Gamma_x \simeq K_p$ for some $p \geqq 0$. The following facts can be proved by induction on the size of the graph involved. For more details about chordal graphs and k-trees see [6], [14], [15], [17].

FACT 3.6. *Every chordal graph which is not complete has two nonadjacent simplicial points.*

FACT 3.7. *Two distinct simplicial points of a proper k-tree T are not adjacent unless T is* K_{k+1}.

FACT 3.8. *Every clique of a proper k-tree has cardinality* $k+1$.

FACT 3.9. *If x is a simplicial point of a proper k-tree T and T is not complete, then* $T\backslash\{x\}$ *is a proper k-tree.*

LEMMA 3.10. *Let C be a chordal graph with maximal clique size* $\leqq k+1$. *Then there exists a proper k-tree T containing C as an induced subgraph such that if C is complete then every vertex of C is a simplicial point of T, and for C noncomplete, at least two vertices of C are simplicial points of T.*

Proof. First note that the lemma is trivially true if C is complete. We will use induction on the cardinality of C. Thus assume that C is not complete and that the lemma holds for proper subgraphs of C. Let x and y be nonadjacent simplicial points of C, and let T' be a proper k-tree containing $C\backslash\{x\}$ as an induced subgraph, satisfying the hypothesis of the lemma. Suppose Γ_x is $\{v_1, \cdots, v_p\}$. Then by Fact 3.8 there exist vertices $v_{p+1}, \cdots, v_{k+1}$ of T' such that $v_1, \cdots, v_{k+1}$ form a clique in T'. Moreover, since y is not adjacent to x, we may assume that if y is one of the vertices $v_1, \cdots, v_{k+1}$, then y is the vertex v_{k+1}. We obtain the proper k-tree T by adjoining vertices $x = x_{p+1}, \cdots, x_{k+1}$ to T' such that x_i is adjacent to v_j for $1 \leqq j \leqq i-1$ and x_i is also adjacent to x_j for $i+1 \leqq j \leqq k+1$. It is easy to see that C is an induced subgraph of T and that x is a simplicial point of T.

We now show that T has another simplicial point which is not adjacent to x. First note that if $C\backslash\{x\}$ is complete, then y is a simplicial point of T'. Since y is adjacent to no vertex in $T\backslash T'$, clearly y is also a simplicial point of T. Thus, we may assume that $C\backslash\{x\}$ has two vertices u and v which are simplicial points of T'. Moreover, by Fact 3.7, u and v are not adjacent, and hence at least one of them, say u, is not v_j for any j. Thus u is a simplicial point of T which is not adjacent to x. □

We are now ready to complete the proof of one direction of Theorem 3.5.

PROPOSITION 3.11. *Suppose* $G = G' \equiv F$, *where* G' *is chordal-clique-complete. Then* $[G, G]$ *has a graph containing a G-bolt.*

Proof. By Lemma 3.4 it suffices to show that $[G', G']$ has a graph containing a G'-bolt. Let $G' = C \equiv K_k$, where C is a noncomplete chordal graph with maximum clique size equal to $k+1$. By Lemma 3.10 and 3.7, we can find a proper k-tree T containing C as an induced subgraph and two vertices x and y of C which are nonadjacent simplicial points of T. Let C' be a graph isomorphic to C. Then, since $T \in [K_k, K_k]$, as noted before, it is easy to see that $T \equiv C' \in [G', G']$. Moreover, as $\Gamma_x(T) \simeq \Gamma_y(T) \simeq K_k$, we have $\Gamma_x(T \equiv C') \simeq \Gamma_y(T \equiv C') \simeq G'$. Choose a subgraph L of C' with $L \simeq K_k$, and let $Z = C \equiv L$. Now it is easy to check that $(x, y, \Gamma_x(T \equiv C'), \Gamma_y(T \equiv C'), Z)$ forms a G'-bolt in $T \equiv C'$. □

We now concentrate on proving the opposite direction of Theorem 3.5. A vertex x of a graph D is said to be a *superpoint* of D if x is adjacent to every other vertex of D. We denote the set of superpoints on D by $s(D)$.

LEMMA 3.12. *If* $[G, G]$ *has a graph which contains a G-bolt, then there exist* $(H, B) \in [G, G]_b$ *and* $z \in H\backslash B$ *such that* $G(z)\backslash(B \cup s(G(z))) \neq \varnothing$, *where* $G(z)$ *is the support of z in* (H, B) (*see beginning of* § 2 *for definition*).

Proof. Let $H' \in [G, G]$ contain a G-bolt (x, y, X, Y, Z). Form H by adding new vertices u, v, z to H' so that $\Gamma_u = X \cup \{z\}$, $\Gamma_v = Y \cup \{z\}$ and $\Gamma_z = (Z \backslash \{x, y\}) \cup \{u, v\}$. Then for any base B of H' it is simple to check that (H, B) is in $[G, G]_b$. Moreover, $z \in H \backslash B$, and since u and v are not adjacent, they both belong to $G(z) \backslash (B \cup s(G(z)))$. □

If U and V are subgraphs of a graph, then we will say that U is completely connected to V if every vertex of U is adjacent to every vertex of V, i.e., $U \cup V$ induces a subgraph isomorphic to $U \equiv V$. The following proposition completes the proof of Theorem 3.5.

PROPOSITION 3.13. *If $[G, G]$ has a graph which contains a G-bolt, then G is chordal-clique-complete-extended.*

Proof. Suppose $[G, G]$ has a graph containing a G-bolt. By Lemma 3.12 there is some $(H, B) \in [G, G]_b$, with $z \in H \backslash B$ and $G(z) \backslash (B \cup s(G(z))) \neq \varnothing$, such that $|H|$ is minimal with respect to this property. By the minimality of $|H|$, it is obvious that for each $y \in G(z) \backslash B$ we must have $(G(y) \backslash B) \subset s(G(y))$. Let A be the union of the sets $B \backslash G(y)$ for $y \in G(z) \backslash B$, let $J = (G(z) \backslash B) \cup (G(z) \cap A)$ and let $C = J \backslash s(J)$. We will prove the following facts about these subgraphs.

(a) $A \subset s(B)$.

(b) C is chordal.

(c) C is nonempty and is not a complete graph.

(d) $|s(G(z))| \geqq q$, where $q+1$ is the size of the largest clique in C.

(e) C is completely connected to $G(z) \backslash C$.

It is clear that $C \cap s(G(z)) = \varnothing$, since by the definition of C we have $C \subset G(z)$ and $s(C) = \varnothing$. Combining this with (b), (c), (d) and (e), it is obvious that G is chordal-clique-complete-extended since if K is any subgraph of $s(G(z))$ of size q, then $G(z) = (C \equiv K) \equiv (G(z) \backslash (C \cup K))$ and $C \equiv K$ is chordal-clique-complete. Thus all that remains is the proof of statements (a) to (e).

Proof of (a). Let $d = |s(G(z))|$. Notice that since $G \simeq B \simeq G(z) \simeq G(y)$ for each y in $G(z) \backslash B$, the number of superpoints in each of these graphs is the same, namely d. Since $(G(y) \backslash B) \subset s(G(y))$, we have $d = |G(y) \backslash B| + |s(B \cap G(y))|$. Clearly, $s(B) \cap (G(y) \cap B) \subset s(B \cap G(y))$ so

$$\begin{aligned} |s(B) \cap (B \backslash G(y))| &= d - |s(B) \cap (B \cap G(y))| \\ &\geqq d - |s(B \cap G(y))| = |G(y) \backslash B| = |B \backslash G(y)|. \end{aligned}$$

This shows that $B \backslash G(y) \subset s(B)$ for each $y \in G(z) \backslash B$.

Proof of (b). Let D be a cycle of length at least four in C. We will show that D has a chord. By (a) we know that A is complete, so this is obvious if $D \subset A$. Thus, we can choose $y \in (G(z) \backslash B) \cap D$ so that $l(y)$ is maximal, where l is the level function with respect to B (defined in § 2). Let x and w be the neighbors of y in D. Then x and w are adjacent. This is obvious if both x and w are in A, so suppose $x \in G(z) \backslash B$. Since x and w are adjacent to y, by the maximality of $l(y)$ we have $x \in G(y) \backslash B$ and $w \in G(y)$; but now x is adjacent to w since $(G(y) \backslash B) \subset s(G(y))$.

Proof of (c). Since we have already noted that $s(C) = \varnothing$, it suffices to show that C is nonempty, and thus, since $C = J \backslash s(J)$ we need only show that J has two nonadjacent vertices. Because $G(z) \backslash (B \cup s(G(z))) \neq \varnothing$, there are nonadjacent vertices u and v with $u \in G(z) \backslash B$ and $v \in G(z)$. Moreover, v is not in $G(u)$ since v is not adjacent to u. Hence, either $v \in G(z) \cap (B \backslash G(u))$ or $v \in G(z) \backslash B$, and in either case, we have both u and v in J.

Proof of (*d*). Let K be any largest clique in C. By (a) it suffices to show that $|A| \geqq |K| - 1$. This is obviously true if $K \subset A$ so we may choose $y \in K \backslash A$ with $l(y)$ maximal. Now $|B \backslash G(y)| = |G(y) \backslash B| \geqq |K \backslash A| - 1$ since the maximality of $l(y)$ implies that $K \backslash \{y\} \subset G(y)$ and the definition of C implies that $(K \backslash A) \subset (K \backslash B)$. Moreover, since $K \backslash \{y\} \subset G(y)$, we have $K \cap (B \backslash G(y)) = \varnothing$, and hence we obtain $|A| \geqq |K \cap A| + |B \backslash G(y)| \geqq (|K| - K \backslash A|) + (|K \backslash A| - 1) = |K| - 1$.

Proof of (e). First notice that $G(z) \backslash B$ is completely connected to $B \backslash A$ since for each $y \in G(z) \backslash B$ we have $B \backslash A \subset (B \cap G(y)) \subset \Gamma_y$. By (a), A is also completely connected to $B \backslash A$, and combining these we see that J and hence C, is completely connected to $B \backslash A$. It is easy to check that $G(z) \backslash C \subset (B \backslash A) \cup s(J)$, and hence C is completely connected to $G(z) \backslash C$, since obviously C is completely connected to $s(J)$. □

4. Polynomial isomorphism for some hookup classes. The results of the preceding section naturally beg the question of the complexity of isomorphism testing in hookup classes $[A, G]$ in which no graph contains a G-bolt. We now give a polynomial algorithm for isomorphism testing in such classes. We begin with two preliminary lemmas which show that if $H \in [A, G]$ does not contain a G-bolt then H has a rather tree-like structure.

Given a base B of a graph $H \in [A, G]$ with B-decomposition $\{B(0), B(1), \cdots, B(p)\}$, recall that the level function is defined on H by $l(x) = i$ if $x \in B(i)$.

LEMMA 4.1. *If $H \in [A, G]$ does not contain a G-bolt, then for any base B of H with B-decomposition $\{B(0), B(1), \cdots, B(p)\}$ we have $|G(z) \cap B(l(z) - 1)| = 1$ for all z with $l(z) \geqq 2$.*

Proof. First note that since $z \in B(l(z))$ we must have $|G(z) \cap B(l(z) - 1)| \geqq 1$ for any z with $l(z) \geqq 1$. Now suppose there exists $z \in H$ with $l(z) \geqq 2$ and $|G(z) \cap B(l(z) - 1)| \geqq 2$. Choose x and y to be distinct vertices of $G(z) \cap B(l(z) - 1)$. Let $X = G(x)$, $Y = G(y)$ and $Z = G(z)$. Then (x, y, X, Y, Z) is a G-bolt in H. To see this note that conditions (3.1.2) and (3.1.3) are trivially satisfied and that (3.1.1) is satisfied due to Lemma 2.1(e). Finally, since $\Gamma_x \cap \bigcup_{j=0}^{l(z)-1} B(j) = G(x)$ by Lemma 2.1(f) and $G(z) \subset \bigcup_{j=0}^{l(z)-1} B(j)$, we have $\Gamma_x \cap G(z) = G(x) \cap G(z)$, and similarly, $\Gamma_y \cap G(z) = G(y) \cap G(z)$, satisfying (3.1.4). Thus we have shown that H has a G-bolt, contradicting the hypothesis. □

Let $(H, B) \in [A, G]_b$ such that H does not contain a G-bolt. For $z \in H$ with $l(z) \geqq 2$, we let $f(z)$ denote the vertex in $G(z) \cap B(l(z) - 1)$. Loosely, $f(z)$ can be thought of as the father of z. We also define $F(z)$ for $z \in H$ with $l(z) \geqq 1$ to be a set of vertices of H as follows:

$$F(z) = \begin{cases} B & \text{if } l(z) = 1, \\ \{f(z)\} \cup F(f(z)) & \text{if } l(z) \geqq 2. \end{cases}$$

Let $T(H)$ be the rooted tree with vertices $\{r\} \cup (H \backslash B)$. The vertex r is the root of $T(H)$ and is father of all vertices in $B(1)$; for any other vertex z in $T(H) \backslash (B(1) \cup \{r\})$, its father is $f(z)$. Now if we identify r with B, it is easy to see that $F(z)$ is the set of vertices, excluding z, on the path from z to r.

LEMMA 4.2. *Let $(H, B) \in [A, G]_b$ such that H does not contain a G-bolt. Then for all z with $l(z) \geqq 1$, we have $G(z) \subset F(z)$.*

Proof. Suppose not. Choose $z \in H$ such that $G(z) \backslash F(z) \neq \varnothing$ and such that $l(z)$ is minimal. Note that $l(z) \geqq 2$ since by definition for $z \in B(1)$ we have $G(z) \subset B = F(z)$. Choose $x \in G(z) \backslash F(z)$ such that $l(x)$ is maximal, and let $y = f(z)$. Let $X = G(x)$,

$Y = G(y)$ and $Z = G(z)$. As before, we will show that (x, y, X, Y, Z) forms a G-bolt in H. Again, (3.1.2) and (3.1.3) are trivially satisfied and likewise $\Gamma_y \cap G(z) = G(y) \cap G(z)$. We also have $\Gamma_x \cap (\bigcup_{j=l(x)+1}^{l(z)-1} B(j)) \cap G(z) = \varnothing$. To see this suppose $w \in \Gamma_x \cap (\bigcup_{j=l(x)+1}^{l(z)-1} B(j)) \cap G(z)$. Then since w and x are adjacent and $l(x) < l(w)$, obviously, $x \in G(w)$. However, as $w \in G(z)$ and $l(w) > l(x)$ by the maximality of $l(x)$, we must have $w \in F(z)$ and hence $F(w) \subset F(z)$. But now this shows that $x \in G(w) \backslash F(w)$, which is impossible by the minimality of $l(z)$ since $l(w) < l(z)$. Both (3.1.1) and the rest of (3.1.4) follow immediately from $\Gamma_x \cap (\bigcup_{j=l(x)+1}^{l(z)-1} B(j)) \cap G(z) = \varnothing$, since this shows that x and y are not adjacent and that $\Gamma_x \cap G(z) = \Gamma_x \cap (\bigcup_{j=0}^{l(x)} B(j)) \cap G(z) = G(x) \cap G(z)$. □

Notice that if we interpret Lemma 4.2 in the context of $T(H)$ it states that for $1 \leqq l(x) \leqq l(y)$ then x adjacent to y implies that x is on the path from y to the root r. We are now ready to present the polynomial isomorphism testing algorithm.

THEOREM 4.3. *If no graph in $[A, G]$ contains a G-bolt, then isomorphism of graphs in $[A, G]$ can be tested in polynomial time with exponent $|A|+2$.*

Proof. We claim that given two graphs H and H' in $[A, G]$ of cardinality n, with bases B and B', we can test whether $(H, B) \simeq (H', B')$ in $O(n^2)$ time.

Using this we can test for isomorphism in $[A, G]$ as follows: Let H and H' be two graphs of cardinality n in $[A, G]$.

Step 1. Find a base B of H. This requires at most $O(n^{|A|+1})$ operations since we can use Algorithm 2.2 to test every subgraph of H with cardinality $|A|$ to see whether it is a base of H and each test requires $O(n)$ operations.

Step 2. For each subgraph B' of H' with cardinality $|A|$, test whether B' is a base of H', and if so test whether $(H, B) \simeq (H', B')$. By the claim above and Algorithm 2.2, this requires at most $O(n^{|A|+2})$ operations. Furthermore, it is clear that $H \simeq H'$ if and only if for some base B' of H' we have $(H, B) \simeq (H', B')$.

We now prove our claim. Let the vertices of B be $v_1, \cdots, v_q$ and those of B' be $v'_1, \cdots, v'_q$. The procedure we describe actually tests whether there exists an isomorphism $\psi: H \to H'$ such that $\psi(v_j) = v'_j$ for each j. Thus to determine whether $(H, B) \simeq (H', B')$ it may be necessary to test all possible $q!$ ways of labelling the vertices of B'.

Let T be the rooted labelled tree formed by labelling the nodes of the rooted tree $T(H)$ as follows. For any vertex x of $T(H)$ other than the root r, let x have the label

$$\{j: v_j \in G(x)\} \cup \{n + j: G(x) \cap B(j) \neq \varnothing, j \geqq 1\}.$$

As usual, $B(0), B(1), \cdots, B(p)$ denotes the B-decomposition of H, $G(x)$ is defined as $\Gamma_x \cap \bigcup_{j=0}^{l(x)-1} B(j)$ and $l(x)$ is the level function of H with respect to B. Similarly, let $B'(0), B'(1), \cdots, B'(s)$ denote the B'-decomposition of H', let $G'(y)$ denote $\Gamma_y \cap \bigcup_{j=0}^{l'(y)-1} B'(j)$ for $y \in H' \backslash B'$, where $l'(y)$ is the level function of H' with respect to B'. Note that we may assume $s = p$ since otherwise obviously $(H, B) \neq (H', B')$. Now let T' denote the rooted labelled tree formed by analogously labelling the nodes of $T(H')$. Thus if y is a nonroot vertex of $T(H')$ then y has the label

$$\{j: v'_j \in G'(y)\} \cup \{n + j: G'(y) \cap B'(j) \neq \varnothing, j \geqq 1\}.$$

Now we claim that there exists an isomorphism $\psi: H \to H'$ such that $\psi(v_j) = v'_j$ if and only if the rooted labelled trees T and T' are isomorphic, and the correspondence $v_j \mapsto v'_j$ induces an isomorphism from the graph B to B'. It is immediate that such an isomorphism ψ yields an isomorphism from T to T', so we will concentrate on proving the opposite direction of this claim.

Suppose $v_j \mapsto v'_j$ is an isomorphism from B to B', and let τ' be an isomorphism from T to T' which preserves both the rooting and labelling. Define a map τ from H to H' by $\tau(x) = \tau'(x)$ if $x \notin B$, and $\tau(v_j) = v'_j$. We will show that τ is an isomorphism.

First note that τ preserves the level function since $\tau(B) = B'$ and if $l(x) \geqq 1$, then $l(x)$ is simply the distance from x to the root in T, whereas $l'(\tau(x))$ is the distance from $\tau(x)$ to the root in T' and hence, $l(x) = l'(\tau(x))$. Suppose x and y are adjacent vertices of H with $1 \leqq l(x) < l(y)$. Then $n + l(x)$ is in the label of y in T so $n + l'(\tau(x))$ is in the label of $\tau(y)$. By Lemma 4.2 we have that $\tau(y)$ is adjacent to the vertex w on the path from $\tau(y)$ to the root of T' such that $l'(w) = l(x)$. However, since by Lemma 4.2 we also have that x is the vertex on the path from y to the root of T with level $l(x)$, we see that w must be $\tau(x)$, and we have shown that $\tau(x)$ and $\tau(y)$ are adjacent. A similarly straightforward argument handles the cases that $l(x) = 0 = l(y)$ and $l(x) = 0 < l(y)$. It is clear that the symmetric argument shows that $\tau(x)$ adjacent to $\tau(y)$ implies that x is adjacent to y, and hence τ is an isomorphism.

The final point to note is that T and T' can be constructed and tested for isomorphism in $O(n^2)$ operations, since the number of elements in each label is $O(|G|)$, all elements of labels are in the set $\{1, 2, \cdots, 2n-1\}$ and the sets $B(j)$, $B'(j)$, $G(x)$, $G'(y)$ for $1 \leqq j \leqq p$, $x \in H \backslash B$, $y \in H' \backslash B'$, are available in $O(n)$ operations by Algorithm 2.2. □

COROLLARY 4.4. *For any graph A and integer $k \geqq 1$, isomorphism in $[A, K_k]$ can be tested in polynomial time. In particular, isomorphism testing of k-trees is polynomial for any given k.*

Proof. No graph can contain a K_k-bolt since every pair of vertices in K_k is adjacent, and hence condition (3.1.3) cannot be satisfied. □

It is, however, interesting to note the following fact.

THEOREM 4.5. *The class of all k-trees, i.e., $\bigcup_{k=1}^{\infty} [K_k, K_k]$, is isomorphism complete.*

Proof. We show how to represent uniquely (and reconstructibly) any graph G of cardinality n as an n-tree $k(G)$, such that $|k(G)| = O(|G|^2)$. Let K be a complete graph with vertices $v_1, \cdots, v_n$, and let the vertices of G be labelled $1, 2, \cdots, n$. Form $k(G)$ as follows. First adjoin a vertex z to K, making z adjacent to every vertex of K. Next, for $i = 1, \cdots, n$, add x_i adjacent to $\{v_j : j \neq i\} \cup \{z\}$. Finally for each pair (i, j) such that i is adjacent to j in G, add $y(i, j)$ adjacent to $\{x_i\} \cup \{v_k : k \neq i, j\} \cup \{z\}$. Obviously $k(G)$ is constructible in $O(|G|^2)$ operations, and $k(G) \in [K_n, K_n]$. □

5. Generalizations. In this section we will consider some possible generalizations of hookup classes. We begin with the concept of a partial hookup class. A graph F is *partially isomorphic* to G, written $F \sim G$, if a partial subgraph on all the vertices of F is isomorphic to G. Equivalently, $F \sim G$ if it is possible to obtain a graph isomorphic to G by removing some of the edges from F. We define the partial hookup class p-$[A, G]$ recursively by $H \in p$-$[A, G]$ if either H is isomorphic to A, or there exists $z \in H$ such that $\Gamma_z \sim G$ and $H \backslash \{z\} \in p$-$[A, G]$. Analogously, a *partial G-bolt* in a graph H is a 5-tuple (x, y, X, Y, Z) such that X, Y, Z are subgraphs of H which are partially isomorphic to G, and x and y are nonadjacent vertices of Z such that $\Gamma_x \cap Z = (X \backslash \{x\}) \cap Z$ and $\Gamma_y \cap Z = (Y \backslash \{y\}) \cap Z$.

It is easy to see, using an analogous construction to that of Theorem 3.2 that if p-$[A, G]$ has a graph containing a partial G-bolt, then p-$[A, G]$ is isomorphism complete. Combining this with the next theorem shows that virtually all nontrivial partial hookup classes are isomorphism complete.

THEOREM 5.1. *The class p-$[A, G]$ contains a partial G-bolt if and only if G is not complete and A has a subgraph partially isomorphic to G.*

Proof. It is obvious that $p\text{-}[A, K_k]$ cannot contain a partial G-bolt since if $Z \sim K_k$ then $Z \simeq K_k$, and hence cannot contain a pair of nonadjacent vertices. Also, if A has no subgraph which is partially isomorphic to G then $p\text{-}[A, G] = \{A\}$, and obviously A does not contain a partial G-bolt.

Now suppose that A has a subgraph G' which is partially isomorphic to G, and let x' and y' be two vertices of G' which correspond to nonadjacent vertices of G via the partial isomorphism. Form H by adding vertices x and y to A so that they are both adjacent to every vertex of G'. Obviously, $H \in p\text{-}[A, G]$. Moreover, if we let $X = \Gamma_x$, $Y = \Gamma_y$, and $Z = \{x, y\} \cup (G' \backslash \{x', y'\})$, it can easily be checked that (x, y, X, Y, Z) is a partial G-bolt in H. □

COROLLARY 5.2. *If G is not a complete graph and A has a subgraph which is partially isomorphic to G, then $p\text{-}[A, G]$ is isomorphism complete; otherwise isomorphism testing in $p\text{-}[A, G]$ is polynomial.*

Proof. This follows from noticing that if G is complete then $p\text{-}[A, G] = [A, G]$, which was shown to have a polynomial isomorphism algorithm in Corollary 4.4. □

We now consider hookup classes with multiple initial and hooking graphs. Thus, we define $H \in [A_1, \cdots A_r; G_1, \cdots, G_s]$ if either $H \simeq A_i$ for some i, or there exists $z \in H$ such that $\Gamma_z \simeq G_j$ for some j and $H \backslash \{z\} \in [A_1, \cdots A_r; G_1, \cdots, G_s]$. A natural conjecture to make is that isomorphism testing in $[A_1, \cdots A_r; G_1, \cdots, G_2]$ is polynomial if and only if it is polynomial in $[A_i, G_j]$ for each i, j. However, some evidence against this is the following example of graphs A, G_1 and G_2 such that isomorphism testing is polynomial in both $[A, G_1]$ and $[A, G_2]$, yet $[A; G_1, G_2]$ is isomorphism complete.

Example 5.3. Let $A = G_1 = K_1$ (a single node) and $G_2 = P_2$ (the path of length 2). Now $[A, G_1]$ is the set of trees, and $[A, G_2] = \{A\}$, so obviously isomorphism testing is polynomial in both $[A, G_1]$ and $[A, G_2]$. Also, G_2 is chordal-clique-complete since $G_2 \simeq C \equiv K_1$, where C is the (chordal) graph with two isolated points, and hence $[G_2, G_2]$ is isomorphism complete. Finally, since $G_2 \in [A, G_1]$, we have $[G_2, G_2] \subset [A; G_1, G_2]$, which shows that $[A; G_1, G_2]$ is isomorphism complete.

It is, however, true that isomorphism testing is polynomial in $[K_1, \cdots, K_r; K_1, \cdots, K_s]$ for any $r, s \geqq 1$. Let us define a var-k-tree as any graph in $[K_1, \cdots, K_k; K_1, \cdots, K_k]$. Clearly, it suffices to give a polynomial algorithm for isomorphism testing of var-k-trees. As in the algorithm given in the preceding section, the basic strategy is to obtain a tree-like representation of var-k-trees, but unfortunately, the representation is somewhat more complicated in this case. For the sake of brevity, we merely sketch out the procedure of obtaining this representation and how the isomorphism testing can be performed.

LEMMA 5.4. *Every connected induced subgraph of a* var-k-*tree is a* var-k-*tree.*

Proof. This follows immediately from noticing that a var-k-tree is simply a connected chordal graph with maximum clique size $\leqq k+1$. □

LEMMA 5.5. *If H is a* var-k-*tree, then*

(1) *The induced subgraph on the simplicial vertices of H is a disjoint union of cliques, say $C(1), \cdots, C(q)$.*
(2) *For any $x, y \in C(i)$, we have $\Gamma_x \cap (H \backslash C(i)) = \Gamma_y \cap (H \backslash C(i))$.*
(3) *Either $H = C(1)$ or $H \backslash \bigcup_{i=1}^{q} C(i)$ is connected and nonempty.*
(4) *If H is not complete, then $H \backslash \bigcup_{i=1}^{q} C(i)$ is a* var-k-*tree.*

Proof. (1) and (2) follow easily from the definition of simplicial vertex, and (4) follows immediately from (3) and Lemma 5.4. Thus all we will prove is (3). Let $H' = H \backslash \bigcup_{i=1}^{q} C(i)$. Obviously if H' is empty then $H = C(1)$ since H is connected. Thus assume H' is nonempty and disconnected. Choose x and y in different

components of H' such that the distance between them in H is minimal, and let $x = x_1, \cdots, x_t = y$ be a shortest path between them in H. By the minimality of the distance, each x_i must be a simplicial vertex for $2 \leqq i \leqq t-1$, and moreover they must all be in the same $C(k)$ by (1). But now by (2) both x and y must be neighbors of x_2, and hence are adjacent as x_2 is a simplicial vertex, a contradiction. □

We call the cliques $C(1), \cdots, C(q)$ the *simplicial sets* of H. For each simplicial set C of H, we define its *simplicial clique* to be $\{x\} \cup \Gamma_x(H)$, where x is any vertex of C. Notice that by Lemma 5.5(2) this definition is independent of the choice of x. It is easy to see that by Lemma 5.5, in polynomial time we can obtain a simplicial set decomposition $\{C(i, j): 1 \leqq i \leqq r, 1 \leqq j \leqq q_i\}$, with the following properties, where r is some positive integer. Let $H(i) = \cup\{C(t, j): i \leqq t \leqq r, 1 \leqq j \leqq q_t\}$. Then $H(1) = H$, each $H(i)$ is a var-k-tree and $C(i, 1), \cdots, C(i, q_i)$ are the simplicial sets of $H(i)$. Moreover, by Lemma 5.5 it follows that the $C(i, j)$ are disjoint and that $q_r = 1$. For any vertex $x \in H$, we define the *age* of x, denoted by $a(x)$, to be i if $x \in C(i, j)$ for some j. In other words, x is of age i if x is a simplicial vertex of $H(i)$. Define $D(i, j)$ to be the simplicial clique of $C(i, j)$ in $H(i)$. Now for any $D(i, j)$ other than $D(r, 1)$ we define its *father*, $f(D(i, j))$, to be $D(m, n)$ such that $D(i, j) \backslash C(i, j) \subset D(m, n)$ and m is maximal with this property. Figure 2 illustrates these notions. The following two lemmas show that $f(D(i, j))$ is well defined.

LEMMA 5.6. *If $D(i, j) \neq D(r, 1)$, then there exists $D(m, n)$ with $m > i$ such that $D(i, j) \backslash C(i, j) \subset D(m, n)$.*

Proof. Let $m = \min\{a(x): x \in D(i, j) \backslash C(i, j)\}$. Obviously $m > i$ and $D(i, j) \backslash C(i, j) \subset H(m)$. Let $x \in D(i, j) \backslash C(i, j)$ such that $x \in C(m, n)$ for some n. Since

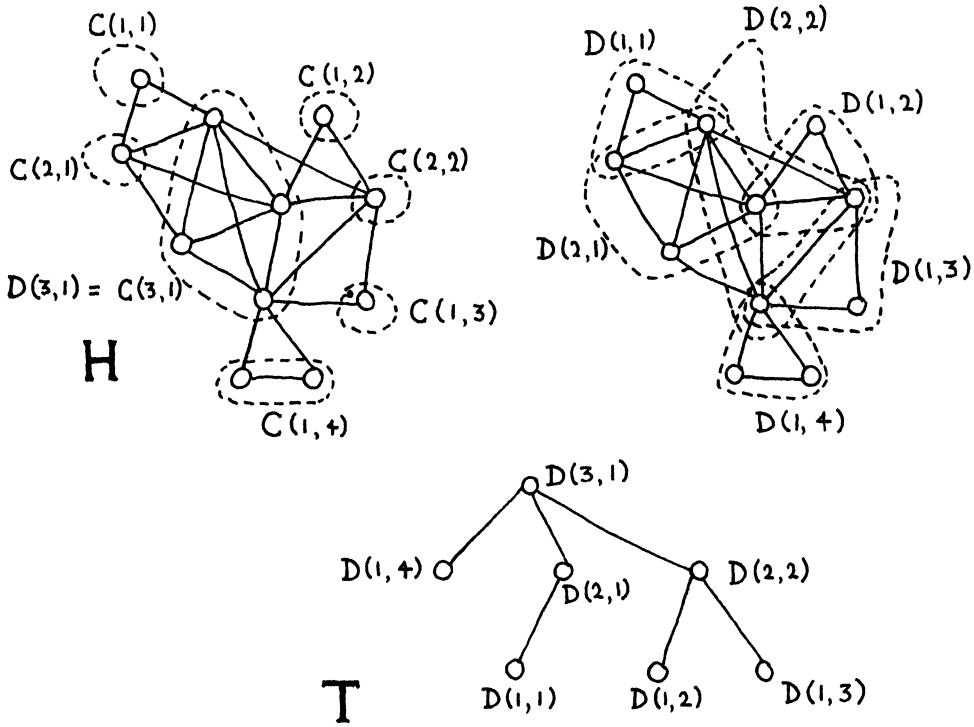


FIG. 2

$D(i,j)\backslash C(i,j)$ is a complete subgraph we must have $D(i,j)\backslash C(i,j) \subset \{x\} \cup \Gamma_x(H(m)) = D(m,n)$. □

LEMMA 5.7. *If $D(i,j)\backslash C(i,j) \subset D(a,b) \cap D(a,c)$ for $b \neq c$ and some $a \geqq i$, then there exists m, n with $m > a$ such that $D(i,j)\backslash C(i,j) \subset D(m,n)$.*

Proof. Since $C(a,b) \cap C(a,c) = \varnothing$, we have $D(i,j)\backslash C(i,j) \subset (D(a,b)\backslash C(a,b))$, and the result follows immediately from Lemma 5.6. □

Thus the father relationship $f(D(i,j))$ defines a rooted tree T on the sets $D(i,j)$ with $D(r,1)$ as the root. We call the set $D(r,1)$ the root-clique.

Let $H(i,j)$ be the induced subgraph of H on the union of the sets $D(m,n)$ such that $D(i,j)$ is on the path from $D(m,n)$ to $D(r,1)$ in T. Thus, in other words, $H(i,j)$ is the subgraph on vertices which are in simplicial cliques in the subtree of T rooted at $D(i,j)$. It is not hard to see that $H(i,j)$ is connected and, hence, is a var-k-tree. Moreover, $D(i,j)$ separates $H(i,j)$ from the rest of H, and the subtree of T rooted at $D(i,j)$ is exactly the tree for $H(i,j)$.

The rest of this section is devoted to developing a polynomial algorithm to determine, given a, b, c with $b \neq c$, whether there exists an isomorphism τ mapping $H(a,b)$ onto $H(a,c)$ such that $\tau(D(a,b)) = D(a,c)$. To see that this yields a polynomial algorithm for determining isomorphism of var-k-trees, note that if F and F' are var-k-trees with root-cliques D and D' then we can form a var-$(k+1)$-tree H by adding a new vertex z to the union of F and F', such that z is adjacent to every vertex in $D \cup D'$. Now F and F' are isomorphic if and only if $\{z\}$ is the root-clique $D(r,1)$ of H, D and D' are the two sons $D(r-1,1)$ and $D(r-1,2)$ of $D(r,1)$ in T, and there is an isomorphism τ mapping $H(r-1,1)$ onto $H(r-1,2)$ such that $\tau(D(r-1,1)) = D(r-1,2)$.

For $1 \leqq a \leqq r-1$ and $1 \leqq b \leqq c \leqq q_a$, we define the mapping set $M(a,b,c)$ to be the set of isomorphisms from $D(a,b)$ to $D(a,c)$ which are extendable to isomorphisms from $H(a,b)$ to $H(a,c)$. We will give an $O((k+1)!|H|)$ algorithm which determines $M(a,b,c)$ given that $M(d,e,f)$ is known for every $d < a$ and $1 \leqq e \leqq f \leqq q_d$. Clearly this yields an $O((k+1)!|H|^3)$ algorithm for determining all the sets $M(a,b,c)$ since T has at most $|H|$ nodes, and as a result, we have an $O(((k+1)!)^2|H|^3)$ algorithm for determining isomorphism of var-k-trees.

Given a one-to-one onto map $\pi: D(a,b) \to D(a,c)$, we say that a son $D(d,e)$ of $D(a,b)$ is π-isomorphic to a son $D(d,f)$ of $D(a,c)$ if $\pi(D(d,e) \cap D(a,b)) = (D(d,f) \cap D(a,c))$ and there is an isomorphism $\tau: H(d,e) \to H(d,f)$ such that τ agrees with π on $D(d,e) \cap D(a,b)$. Clearly such an isomorphism exists if and only if there exists $\theta \in M(d,e,f)$ such that $\theta = \pi$ on $D(d,e) \cap D(a,b)$. Note that for any pair of sons $D(d,e)$ and $D(g,h)$ of $D(a,b)$, we have $H(d,e) \cap H(g,h) \subset D(a,b)$. This ensures that $\pi \in M(a,b,c)$ if and only if we can find a one-to-one correspondence σ between the sons of $D(a,b)$ and the sons of $D(a,c)$, such that $D(d,e)$ and $\sigma(D(d,e))$ are π-isomorphic for each son $D(d,e)$ of $D(a,b)$. Moreover, the same fact shows that if $D(d,e)$ is π-isomorphic to $D(d,f)$ then such a correspondence σ exists if and only if there is an analogous correspondence between $\{D: D$ is a son of $D(a,b)$ and $D \neq D(d,e)\}$ and $\{D: D$ is the son of $D(a,c)$ and $D \neq D(d,f)\}$. Gathering all these facts together, we see that the following algorithm determines $M(a,b,c)$ in $O((k+1)!|H|)$ time, given that $M(d,e,f)$ is known for all $d < a$.

ALGORITHM 5.8.

INPUT: A var-k-tree with its tree representation, a level a, mapping sets $M(d,e,f)$ for each $d < a$, $1 \leqq e, f \leqq q_d$, and simplicial cliques $D(a,b)$ and $D(a,c)$.

OUTPUT: The mapping set $M(a, b, c)$ of isomorphisms from $D(a, b)$ to $D(a, c)$ which are extendable to isomorphisms from $H(a, b)$ to $H(a, c)$.
IF $|D(a, b)| \neq |D(a, c)|$ THEN $M(a, b, c) = \varnothing$;
ELSE for each one-to-one onto map $\pi: D(a, b) \to D(a, c)$ DO;
  YES := TRUE;
  Let all sons of $D(a, b)$ be unmarked;
  For each son D of $D(a, c)$ DO;
    IF there is an unmarked son of $D(a, b)$ which is π-isomorphic to D
      THEN mark it;
    ELSE YES := FALSE;
    END;
IF any son of $D(a, b)$ is still unmarked THEN YES := FALSE;
IF YES = TRUE THEN add π to $M(a, b, c)$:
END;

Acknowledgment. We would like to thank the referee for suggesting Example 5.3, which is much simpler than our original example, and also for several other helpful comments.

REFERENCES

[1] K. S. BOOTH AND C. J. COLBOURN, *Problems polynomially equivalent to graph isomorphism*, T.R.C.S. 77-04, Dept. of Computer Science, Univ. of Waterloo, Canada, 1979.
[2] D. G. CORNEIL, *Recent results on the graph isomorphism problem*, Proc. 8th Manitoba Conference on Numerical Math. and Computing, 1978, pp. 13–31.
[3] A. K. DEWDNEY, *Extensions and generalizations of graph theorems to complexes and hypergraphs*, Ph.D. Thesis, Univ. of Waterloo, Ontario, Canada, 1974.
[4] A. FARLEY, *Networks immune to isolated failures*, CS-TR-79-21, Dept. of Comp. and Inf. Sci., Univ. of Oregon, Eugene, 1979, Networks, 11 (1981), pp. 255–268.
[5] I. S. FILOTTI AND J. N. MAYER, *A polynomial-time algorithm for determining the isomorphism of graphs of fixed genus*, Proc. 12th Annual ACM Symposium on Theory of Computing, 1980, pp. 236–243.
[6] D. R. FULKERSON AND O. A. GROSS, *Incidence matrices and interval graphs*, Pacific H. Math., 15 (1965), pp. 835–855.
[7] S. HEDETNIEMI, *Characterizations and constructions of minimally 2-connected graphs and minimally strong digraphs*, Proc. 2nd Louisiana Conference on Combinatorics, Graph Theory, and Computing, 1971, pp. 257–282.
[8] S. HEDETNIEMI, Private communications, 1977.
[9] J. E. HOPCROFT AND R. E. TARJAN, *Isomorphism of planar graphs*, Complexity of Computer Computations, R. E. Miller and J. W. Thatcher, eds., Plenum Press, New York, 1972, pp. 131–152.
[10] J. E. HOPCROFT AND J. K. WONG, *Linear time algorithm for isomorphism of planar graphs*, Proc. 6th Annual ACM Symposium on Theory of Computing, 1974, pp. 172–184.
[11] E. LUKS, *Isomorphism of bounded valence can be tested in polynomial time*, Proc. 21st Annual IEEE Symposium on Foundations of Computer Science, 1980, pp. 42–49.
[12] G. MILLER, *Isomorphism testing for graphs of bounded genus*, Proc. 12th Annual ACM Symposium on Theory of Computing, 1980, pp. 226–235.
[13] S. L. MITCHELL, *Algorithms on trees and maximal outer-planar graphs: design complexity, analysis and data-structures studies*, Ph.D. Thesis, Univ. of Virginia, Charlottesville 1976.
[14] A. PROSKUROWSKI, *Centres of k-trees*, CS-TR-79-9, Dept. of Computing and Information Sciences, Univ. of Oregon, Eugene, 1979.
[15] ———, *Separating subgraphs in cables and caterpillars*, CS-TR-79-18, Dept. of Comp. and Inf. Sci., Univ. of Oregon, Eugene, 1979.
[16] R. C. READ AND D. G. CORNEIL, *The graph isomorphism disease*, J. Graph Theory, 1 (1977) pp. 339–363.
[17] D. J. ROSE, *On simple characterizations of k-trees*, Discrete Math., 7 (1974), pp. 317–322.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 2, June 1982

0196-5212/82/0302-0013 $01.00/0

REPRESENTATIONS OF $\mathfrak{sl}(2, \mathbb{C})$ ON POSETS AND THE SPERNER PROPERTY*

ROBERT A. PROCTOR†

Abstract. A ranked partially ordered set is said to be *Sperner* if it has no antichain bigger than its largest rank. A necessary and sufficient condition for a ranked partially ordered set to be rank symmetric, rank unimodal and strongly Sperner is presented. This condition involves representations of $\mathfrak{sl}(2, \mathbb{C})$. It is used to provide a new, short proof that this combination of properties is preserved under the product operation. The sufficient part of this condition is also used to provide new, simpler proofs that certain combinatorially interesting partially ordered sets are rank symmetric, rank unimodal and strongly Sperner.

1. Introduction.

DEFINITION. A *ranked poset P of length r* is a partially ordered set P together with a partition $P = \bigcup_{i=0}^{r} P_i$ into $r+1$ ranks P_i, $0 \leqq i \leqq r$, such that elements in P_i cover only elements in P_{i-1}.

DEFINITIONS. A ranked poset P is *Sperner* if no antichain has more elements than the largest rank of P does. It is *strongly Sperner* if for every $k \geqq 1$ no union of k antichains contains more elements than the union of the k largest ranks of P does.

DEFINITIONS. A ranked poset of length r is *rank symmetric* if $|P_i| = |P_{r-i}|$ for $0 \leqq i < r/2$. It is *rank unimodal* if $|P_0| \leqq |P_1| \leqq \cdots \leqq |P_k| \geqq |P_{k+1}| \geqq \cdots \geqq |P_r|$ for some $0 \leqq k \leqq r$.

We follow [PSS] in using the following terminology:

DEFINITION. A ranked poset is *Peck* if it is rank symmetric, rank unimodal and strongly Sperner.

The main result of this paper is a new necessary and sufficient condition for a ranked poset to be Peck. This condition combines a linear algebra/combinatorial lemma of Stanley [Sta] with a technique which uses representations of $\mathfrak{sl}(2, \mathbb{C})$. We will refer to this condition as the *representation condition*. Each aspect of the representation theory of $\mathfrak{sl}(2, \mathbb{C})$ used will be stated clearly for the benefit of readers who are unfamiliar with the subject.

The necessity and the sufficiency of the representation condition are combined in § 4 to produce a new "one line" proof that the product of Peck posets is Peck. Although the statement of this result is purely combinatorial, no nonalgebraic proof is known.

Stanley originally used the lemma mentioned above in conjunction with some techniques from algebraic geometry to prove that a certain collection of posets arising in algebraic geometry were Peck [Sta]. After noting that the representation condition can be applied in principle to all of these posets, we will explicitly describe how to apply this condition to the most combinatorially interesting cases. Two problems in traditional combinatorics can be solved using the result that these cases are Peck [Sta]. Both the proof of this method and its application to the cases considered are somewhat more elementary than Stanley's original treatment. Although we have not found any interesting posets besides those considered by Stanley to which to apply the representation condition, it is formulated in the context of arbitrary ranked posets and is not restricted to posets arising from algebraic geometry.

* Received by the editors August 11, 1981, and in final form August 20, 1981. This research was supported in part by a National Science Foundation Graduate Fellowship and by the Mittag-Leffler Institute.

† Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. Current address: Department of Mathematics, University of California, Los Angeles, Los Angeles, California 90024.

It is possible to translate the Lie representation theory used in this paper to mildly complicated linear algebra. However, this does not simplify the content of the proof in any essential way. Such a translation of the sufficient part of the representation condition in two special cases is presented in the expository article [Pr1].

Readers interested in this technique of using linear algebra in extremal order theory should be aware of papers by Saks [Sak] and Gansner [Gan]. The Jordan canonical form viewpoint presented in these papers led to the discovery of the necessity of the representation condition.

2. Main result. Associate to any ranked poset

$$P = \bigcup_{i=0}^{r} P_i$$

a graded complex vector space

$$\tilde{P} = \bigoplus_{i=0}^{r} \tilde{P}_i,$$

where $\tilde{P}_i$ is the complex vector space freely generated by vectors $\tilde{a}$ corresponding to elements of P_i. A linear operator X on $\tilde{P}$ is a *lowering operator* if $X\tilde{P}_i \subseteq \tilde{P}_{i-1}$. It is a *raising operator* if $X\tilde{P}_i \subseteq \tilde{P}_{i+1}$. A raising operator defined by

$$X\tilde{a} = \sum \Theta(a, b)\tilde{b}$$

is an *order raising operator* if $\Theta(a, b) \neq 0$ implies b covers a. For any ranked poset P of length r, define a linear operator H on $\tilde{P}$ by

$$H\tilde{a} = (2i - r)\tilde{a}$$

when $a \in P_i$.

The Lie algebra $\mathfrak{sl}(2, \mathbb{C})$ consists of all 2×2 trace zero complex matrices with Lie algebra multiplication given by $[u, v] = uv - vu$. The basis usually taken for $\mathfrak{sl}(2, \mathbb{C})$ is [Hum, p. 31]

$$x = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad y = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad h = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

The relations $[x, y] = h$, $[h, x] = 2x$ and $[h, y] = -2y$ completely describe the algebra structure of $\mathfrak{sl}(2, \mathbb{C})$. A representation of $\mathfrak{sl}(2, \mathbb{C})$ on a complex vector space V can be thought of as a choice of three linear operators X, Y and H on V such that $XY - YX = H$, $HX - XH = 2X$ and $HY - YH = -2Y$. An eigenvector for H with eigenvalue λ is referred to as a "weight vector" of the representation of "weight" λ. Any $(d+1)$-dimensional irreducible representation of $\mathfrak{sl}(2, \mathbb{C})$ has as a basis a "string" of vectors $v_0, v_1, \cdots, v_d$ with $Hv_j = (2j - d)v_j$, $Xv_j = v_{j+1}$ and $Yv_j = j(d - j + 1)v_{j-1}$ [Hum, p. 32].

We now present the *representation condition.*

DEFINITION. Let P be a ranked poset of length r. The poset P *carries a representation of* $\mathfrak{sl}(2, \mathbb{C})$ if there exist a lowering operator Y and an order raising operator X on $\tilde{P}$ such that $XY - YX = H$.

Note that $HX - XH = 2X$ is true for all raising operators X, and $HY - YH = -2Y$ is true for all lowering operators Y. Hence X, Y and H define a representation of $\mathfrak{sl}(2, \mathbb{C})$ on $\tilde{P}$ whenever the requirement of the definition, $XY - YX = H$ is satisfied. If P does carry a representation of $\mathfrak{sl}(2, \mathbb{C})$, then the rank subspace $\tilde{P}_i$ is the weight space of weight $2i - r$ for the representation.

LEMMA (Stanley [Sta]). *A ranked poset P of length r is Peck if and only if there exists an order raising operator X on $\tilde{P}$ such that*

$$X^{r-2i}|_{\tilde{P}_i}: \tilde{P}_i \to \tilde{P}_{r-i}$$

is an isomorphism for every $0 \leqq i < r/2$.

THEOREM 1. *A ranked poset is Peck if and only if it carries a representation of $\mathfrak{sl}(2,\mathbb{C})$.*

Proof. Let P be a ranked poset carrying a representation of $\mathfrak{sl}(2,\mathbb{C})$ with order raising operator X. Use complete reducibility [Hum, p. 28] to express this representation as a direct sum of irreducible representations. Each of the irreducible representations has as a basis a "string" of vectors of the form described above. These strings collectively form a new basis for $\tilde{P}$. If one of the irreducible representations has dimension $d+1$, then exactly one of its $d+1$ basis vectors falls in each of the middle $d+1$ consecutive rank subspaces $\tilde{P}_{(r-d)/2}, \tilde{P}_{(r-d)/2+1}, \cdots, \tilde{P}_{(r+d)/2}$. Now the set of new basis vectors falling in a given rank subspace form a basis for that rank subspace, and a given string has a member falling in $\tilde{P}_{r-i}$ if and only if it also has a member falling in $\tilde{P}_i$. Finally, note that the operator X^{d-2j} is an isomorphism from the jth to the $(d-j)$th weight space in any irreducible $(d+1)$-dimensional representation, where $0 \leqq j < d/2$. Conclude that X^{r-2i} is an isomorphism from $\tilde{P}_i$ to $\tilde{P}_{r-i}$, implying by the lemma that P is Peck.

Conversely, suppose that P is Peck. Using the raising order operator X produced by the lemma, construct a new basis for $\tilde{P}$. Let v be any nonzero vector in $\tilde{P}_0$. Then $v, Xv, X^2v, \cdots, X^rv$ are nonzero and linearly independent. Let V denote the subspace spanned by these vectors. Let s be as small as possible such that $\tilde{P}_s$ is not contained in V. Choose a nonzero w in $\tilde{P}_s$ lying outside $\tilde{P}_s \cap V$ and such that $X^{r-2s+1}w = 0$. (Let $w' \in P_s - P_s \cap V$. If $X^{r-2s+1}w' \neq 0$, then $X^{r-2s+1}w' \in V$. Find $z \in P_s \cap V$ such that $X^{r-2s+1}z = -X^{r-2s+1}w'$. Set $w = w' + z$.) Then $w, Xw, X^2w, \cdots, X^{r-2s}w$ are nonzero and linearly independent, and none lies in V. Let W denote the subspace spanned by all basis vectors generated so far. Repeat this procedure until all of $\tilde{P}$ is spanned. The new basis is a disjoint union of strings of vectors, with each string symmetric about the middle rank subspace of $\tilde{P}$. Define a lowering operator Y on $\tilde{P}$ with respect to the new basis. Let $u, Xu, X^2u, \cdots, X^{r-2t}u$ be a typical string of basis vectors. Set $Y[X^ju] = j(t-j+1)[X^{j-1}u]$. Then X^ju is an eigenvector of the operator $XY - YX$ with eigenvalue $2j-(r-2t)$. If $t+j=i$, then $2j-(r-2t) = 2i-r$. In other words, all new basis vectors lying in $\tilde{P}_i$ are eigenvectors for $XY - YX$ with eigenvalue $2i-r$. Therefore $XY - YX = H$, and P carries a representation of $\mathfrak{sl}(2,\mathbb{C})$.

3. Uniqueness of lowering operator. The following fact will not be used in this paper, but seems worthy of mention.

PROPOSITION 1. *Let P be a ranked poset. Let X be a fixed order raising operator on $\tilde{P}$. Then there is at most one lowering operator Y on $\tilde{P}$ such that P carries a representation of $\mathfrak{sl}(2,\mathbb{C})$. That is, if X is fixed, then Y is unique if it exists.*

The proposition above is actually a restatement in the present context of the proposition below. For a given representation of $\mathfrak{sl}(2,\mathbb{C})$, let X, Y and H denote the images of the usual basis x, y and h.

PROPOSITION 2. *In any representation of $\mathfrak{sl}(2,\mathbb{C})$, the image Y is completely determined by the images X and H.*

By change of basis, any representation of $\mathfrak{sl}(2,\mathbb{C})$ can be put into the form used in the proof of Theorem 1. The matrices representing x, y and h are $\mathscr{X} = X_{d_1} \oplus \cdots \oplus X_{d_k}$, $\mathscr{Y} = Y_{d_1} \oplus \cdots \oplus Y_{d_k}$ and $\mathscr{H} = H_{d_1} \oplus \cdots \oplus H_{d_k}$, where $X_d = E_{0,1} + \cdots + E_{d-1,d}$,

$Y_d = dE_{1,0} + \cdots + j(d-j+1)E_{j,j-1} + \cdots + dE_{d,d-1}$ and $H_d = dE_{0,0} + (d-2)E_{1,1} + \cdots + (-d)E_{d,d}$. (Here $E_{i,j}$ denotes the $(d+1)\times(d+1)$ matrix with (i,j)th entry equal to unity and all other entries equal to zero.)

LEMMA. *Let $\mathscr{X}$, $\mathscr{Y}$ and $\mathscr{H}$ be as above for some values of $k, d_1, \cdots d_k$. Then any matrix which commutes with $\mathscr{X}$ and $\mathscr{H}$ also commutes with $\mathscr{Y}$.*

Proof. Suppose $S\mathscr{X} = \mathscr{X}S$ and $S\mathscr{H} = \mathscr{H}S$. Without loss of generality, assume $k = 2$. Set $m = d_1$, $n = d_2$ and

$$S = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

Then $AX_m = X_m A$ and $AH_m = H_m A$, which implies that A is a scalar matrix. Also, $BX_n = X_m B$ and $BH_n = H_m B$, which implies that B is zero if $m \neq n$ or a scalar if $m = n$. Analogous conclusions can be drawn for D and C. So if $m \neq n$, we are done. But even if $m = n$, it is easily checked that $S\mathscr{Y} = \mathscr{Y}S$.

Proof of Proposition 2. Let matrices X, H and Y define one representation of $\mathfrak{sl}(2, \mathbb{C})$, and let X, H and Y' define another representation. Since both representations have the same character, they are equivalent [Hum, p. 125]. Let R be an invertible matrix such that $RXR^{-1} = X$, $RHR^{-1} = H$ and $RYR^{-1} = Y'$. Let S be an invertible matrix which puts the first representation into the canonical form described above: $SXS^{-1} = \mathscr{X}$, $SHS^{-1} = \mathscr{H}$, and $SYS^{-1} = \mathscr{Y}$. Then $SRS^{-1}SXS^{-1}SRS^{-1} = SXS^{-1}$, i.e., $(SRS^{-1})\mathscr{X}(SRS^{-1}) = \mathscr{X}$, and similarly for $\mathscr{H}$. By the lemma, $(SRS^{-1})\mathscr{Y}(SRS^{-1}) = \mathscr{Y}$, implying $SRYR^{-1}S^{-1} = SYS^{-1}$. Therefore $Y = RYR^{-1} = Y'$, i.e., Y is unique.

4. Products of Peck posets. The following theorem was the main result of [Can] and [PSS]. Linear algebra, viz., Stanley's lemma, was used in both proofs.

THEOREM 2. *The product of Peck posets is Peck.*

Proof. Let P and Q be Peck posets. By Theorem 1, these posets each carry a representation of $\mathfrak{sl}(2, \mathbb{C})$, say with operators X', Y', H' and X'', Y'', H'', respectively. The tensor product $\tilde{P} \otimes \tilde{Q}$ and the vector space $\widetilde{P \times Q}$ are isomorphic as graded vector spaces. Let I' and I'' denote the identity operators on $\tilde{P}$ and $\tilde{Q}$. It is easy to verify by direct computation that the operators $X' \otimes I'' + I' \otimes X''$, $Y' \otimes I'' + I' \otimes Y''$ and $H' \otimes I'' + I' \otimes H''$ define a representation of $\mathfrak{sl}(2, \mathbb{C})$ on $\tilde{P} \otimes \tilde{Q}$. (This is true by inspection to readers who are familiar with the definition of the tensor product of two representations of a Lie algebra [Hum, p. 26].) It is also easy to verify that $Y' \otimes I'' + I' \otimes Y''$ and $X' \otimes I'' + I' \otimes X''$ are lowering and order raising operators respectively for $P \times Q$. Apply Theorem 1 to conclude that $P \times Q$ is Peck.

5. Applications. Bruhat posets (defined on Weyl groups) are a certain set of partially ordered sets arising in algebraic geometry. Stanley originally used the lemma in § 2 in conjunction with the hard Lefschetz theorem of algebraic geometry to show that all Bruhat posets are Peck [Sta]. In this section we explicitly describe representations of $\mathfrak{sl}(2, \mathbb{C})$ on the Bruhat posets which are the most interesting from a combinatorial viewpoint. Application of Theorem 1 then reproduces Stanley's result in these cases. The method used here was developed to avoid the use of algebraic geometry. After this alternative method was developed, it was discovered that some proofs of the hard Lefschetz theorem actually proceed by constructing a representation of $\mathfrak{sl}(2, \mathbb{C})$. Under the identifications made in Stanley's work, it can be seen that this representation meets the requirements of Theorem 1 in the case of the Bruhat orders. This guarantees that Theorem 1 can in principle always be applied directly to any Bruhat order.

DEFINITION. A *uniquely modular poset* is a ranked poset satisfying:

(i) Whenever two elements both cover a third element, then there exists a unique fourth element covering both of them.

(ii) Whenever two elements are both covered by a third element, then there exists a unique fourth element covered by both of them.

Finding suitable operators X and Y for an arbitrary ranked poset requires the solution of a system of quadratic equations. When dealing with uniquely modular posets, it is simpler (but less general) to seek representations of $\mathfrak{sl}(2, \mathbb{C})$ of a certain form. This involves solving a system of linear equations in order to satisfy the requirements in the following definition.

DEFINITION. A uniquely modular poset P of length r is *edge-labelable* if each covering relationship $a < d$ can be assigned a rational number $y(d, a)$ such that:

(i) If d covers both a and b and both a and b cover c, then $y(d, a) = y(b, c)$.

(ii) If $a \in P_i$, then

$$\sum_{a \text{ covers } c} y(a, c) - \sum_{d \text{ covers } a} y(d, a) = 2i - r.$$

Pictorially, each edge of the Hasse diagram of the poset is to be labeled with a rational number such that opposite edges in any "square" must receive the same number and such that, for any element in the ith rank, the sum of the labels of edges emanating below the element minus the sum of the labels of edges emanating above the element must equal $2i - r$.

PROPOSITION 3. *Edge-labelable uniquely modular posets are Peck.*

Proof. Let P be an edge-labelable uniquely modular poset. Define an order raising operator X by

$$X\tilde{a} = \sum_{b \text{ covers } a} \tilde{b}$$

for all $a \in P$ and a lowering operator Y by

$$Y\tilde{b} = \sum_{b \text{ covers } a} y(b, a)\tilde{a}.$$

Now confirm that conditions (i) and (ii) in the definition of edge-labelable imply that $XY - YX = H$, where H is defined as usual. Apply Theorem 1.

Notation. Let $\mathbf{m}$ denote a total order with m elements. For any poset P, let $J(P)$ denote the lattice of order ideals of P.

THEOREM 3. *The distributive lattices* $J(\mathbf{m} \times \mathbf{n})$, $J^2(\mathbf{2} \times \mathbf{n-1})$, $J^n(\mathbf{2} \times \mathbf{2})$, $J^3(\mathbf{2} \times \mathbf{3})$ and $J^4(\mathbf{2} \times \mathbf{3})$, with $m \geqq 0$, $n \geqq 1$, *are edge-labelable and therefore Peck.*

Proof. The lattice $J(\mathbf{m} \times \mathbf{n})$ can be described as the set of n-tuples $(a_1, a_2, \cdots, a_n)$, where $0 \leqq a_1 \leqq a_2 \leqq \cdots \leqq a_n \leqq m$, with order given by $a \leqq b$ if $a_i \leqq b_i$ for all i. If b covers a with $a_i = b_i - 1$, set $y(b, a) = (m + n - a_i - i)(a_i + i)$. The lattice $J^2(\mathbf{2} \times \mathbf{n-1})$ can be described as the set of n-tuples $(a_1, a_2, \cdots, a_n)$, where $0 = a_1 = \cdots = a_k < a_{k+1} < \cdots < a_n \leqq n$, $1 \leqq k \leqq n$, with order given by $a \leqq b$ if $a_i \leqq b_i$ for all i. If b covers a with $a_i = b_i - 1$, set $y(b, a) = n(n+1)/2$ if $a_i = 0$, otherwise set $y(b, a) = n(n+1) - a_i(a_i + 1)$. The edge labels for $J^n(\mathbf{2} \times \mathbf{2})$ are $1(2n+2)$, $2(2n+1)$, $\cdots$, $n(n+3)$, $(n+1)(n+2)/2$ and $(n+1)(n+2)/2$; the labels for $J^3(\mathbf{2} \times \mathbf{3})$ are 16^{12}, 30^{12}, 42^6 and 22^6 (exponents indicate the number of times each occurs); and the labels for $J^4(\mathbf{2} \times \mathbf{3})$ are 27^{12}, 52^{12}, 75^{12}, 96^{12}, 66^{12}, 34^{12} and 49^{12}. The verification of requirement (ii) for $J(\mathbf{m} \times \mathbf{n})$ was performed in [Pr1]. We leave the verification of this requirement in the other cases to the reader.

It is easy to verify directly by inspection that the last three cases of the theorem are Peck. These cases are included for the sake of completeness: In a future paper [Pr3], we will prove that there are no other irreducible (with respect to product) edge-labelable distributive lattices besides those listed in Theorem 3. The representations of $\mathfrak{sl}(2,\mathbb{C})$ described in the proof of this theorem arise in a natural way from minuscule representations of semisimple Lie algebras. They also arise in the context of the Hodge identities on the minuscule flag manifolds (viewed as Kähler manifolds). Both of these connections will be described in [Pr2].

DEFINITION. The *poset of shuffles* on $1^{k_1}2^{k_2}\cdots m^{k_m}$ is the set of all sequences with k_1 1's, k_2 2's, $\cdots$, k_m m's, with order generated by the relations

$$\cdots a_i \cdots a_j \cdots \leqq \cdots a_j \cdots a_i \cdots$$

(ith and jth entries interchanged) when $a_j \leqq a_i$. The unique maximal element is $1\cdots 12\cdots 2\cdots m\cdots m$.

The lattice $J(\mathbf{m}\times\mathbf{n})$ is the shuffle poset $1^m 2^n$. The shuffle posets constitute all Bruhat orders of "type A".

We will say that a poset Q is a *cover suborder* of a poset P if Q and P are partial orders on the same set and the order on Q is generated by a subset of the covering relations of P. Note then that if P and Q are also ranked with the same ranking, then P is Peck if Q is Peck.

THEOREM 4. *The shuffle posets are Peck.*

Proof. Given the shuffle poset on $1^{k_1}2^{k_2}\cdots m^{k_m}$, the poset $J(\mathbf{k}_1\times\mathbf{k}_2)\times J((\mathbf{k}_1+\mathbf{k}_2)\times\mathbf{k}_3)\times\cdots\times J((\mathbf{k}_1+\mathbf{k}_2+\cdots+\mathbf{k}_{m-1})\times\mathbf{k}_m)$ is one of $\binom{2m}{m}/(m+1)$ such easily formed suborders. It is easy to check that the product of edge-labelable posets is edge-labelable. Hence the shuffle poset at hand has a Peck cover suborder with the same ranking and is therefore Peck itself.

REFERENCES

[Can] E. R. CANFIELD, *A Sperner property preserved by products*, Linear and Multilinear Algebra, 9 (1980), pp. 151–157.

[Gan] E. GANSNER, *Acyclic digraphs, Young tableaux and nilpotent matrices*, this Journal, 2 (1981), pp. 429–440.

[Hum] J. E. HUMPHREYS, *Introduction to Lie Algebras and Representation Theory*, Springer-Verlag, New York, 1972.

[Pr1] R. PROCTOR, *Solution of two difficult combinatorial problems with linear algebra*, Amer. Math. Monthly, to appear.

[Pr2] ———, *Bruhat lattices, plane partition generating functions and minuscule representations*, in preparation.

[Pr3] ———, *A Dynkin diagram classification theorem arising from a combinatorial problem*, in preparation.

[PSS] R. PROCTOR, M. SAKS AND D. STURTEVANT, *Product partial orders with the Sperner property*, Discrete Math., 30 (1980), pp. 173–180.

[Sak] M. SAKS, *Dilworth numbers, incidence maps and product partial orders*, this Journal, 1 (1980), pp. 211–215.

[Sta] R. STANLEY, *Weyl groups, the hard Lefschetz theorem and the Sperner property*, this Journal, 1 (1980), pp. 168–184.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 3, September 1982

0196-5212/82/0303-0001 $01.00/0

A CLASS OF PERFECT GRAPHS*

JAMES B. SHEARER†

Abstract. Let P be a simply connected polyomino. Let $G(P)$ be the graph whose vertices are the maximal rectangles in P, two such vertices being adjacent if the corresponding rectangles have nontrivial intersection. In this paper we show that $G(P)$ is perfect. This solves a problem posed by Berge et al.

Let P be a polyomino (i.e., a finite subset of the squares in an infinite checkerboard). Let $G(P)$ be the graph derived from P as follows. Let the vertices of $G(P)$ be the maximal rectangles contained (as subpolyominoes) in P. Let two vertices in $G(P)$ be joined by an edge if the rectangles have a nontrivial (i.e., containing at least one unit square of P) intersection. For any graph G let $\alpha(G)$ denote the independence number of G (i.e., the maximum cardinality of a set of nonadjacent vertices of G) and let $\theta(G)$ denote the clique covering number of G (i.e., the minimum cardinality of a collection of cliques in G containing every vertex of G). Clearly $\theta(G) \geqq \alpha(G)$. We say G is perfect whenever $\alpha(G') = \theta(G')$ for all induced subgraphs G' of G. In this paper we prove the following theorem (thereby solving a problem posed in [1]).

THEOREM 1. *Let P be simply connected. Then $G(P)$ is a perfect graph.*

Before proving Theorem 1 we mention the following consequence. Let P be a simply connected polyomino. Let $\alpha(P)$ be the maximum cardinality of a collection of disjoint maximal rectangles in P. Let $\theta(P)$ be the minimum cardinality of a collection of unit squares in P with the property that every maximal rectangle in P contains at least one square of the collection. Then $\alpha(P) = \theta(P)$. This follows at once from Theorem 1 and the identities $\alpha(P) = \alpha(G(P))$ and $\theta(P) = \theta(G(P))$. The first identity is trivial and the second depends on the following fact which is easy to prove.

FACT 1. *If C is a clique in $G(P)$ then P contains a unit square which is contained in all rectangles in C.*

Berge et al. in [1] give an example which shows that when P is not simply connected $\alpha(P)$ need not equal $\theta(P)$.

We now give the proof of Theorem 1. We will show $\alpha(G') = \theta(G')$ for all induced subgraphs G' of $G(P)$ by induction on the number of vertices in G'. Clearly $\alpha(G') = \theta(G') = 1$ if G' consists of a single vertex. Next let G' be an induced subgraph of $G(P)$ on $n > 1$ vertices. By the induction hypothesis $\alpha(G'') = \theta(G'')$ for all proper induced subgraphs G'' of G'. We wish to show $\alpha(G') = \theta(G')$. Clearly we may assume G' is connected. The following four lemmas proved under the above assumptions are in or follow easily from the literature on minimal imperfect graphs (see, for instance, [2]). However, we give proofs for completeness.

LEMMA 1. *Suppose G' contains a clique C whose removal disconnects G' (i.e., the vertex set of G' can be partitioned into 3 nonempty classes A, B and C such that the subgraph of G' induced by C is a clique and G' contains no edges between points in A and points in B). Then $\alpha(G') = \theta(G')$.*

Proof. Let $G'(S)$ denote the subgraph of G' induced by S where S is a subset of the vertex set of G'. Let $r_1 = \alpha(G'(A)) = \theta(G'(A))$, $r_2 = \alpha(G'(B)) = \theta(G'(B))$. Clearly $r_1 + r_2 \leqq \alpha(G') \leqq r_1 + r_2 + 1$. Suppose C contains a point c such that $\alpha(G'(A \cup c)) = r_1 + 1$ and $\alpha(G'(B \cup c)) = r_2 + 1$. Then $\alpha(G'(A \cup B \cup c)) = r_1 + r_2 + 1$. Hence we must have

* Received by the editors June 29, 1981. This research was supported in part by the Office of Naval Research under grant N00014-76-C-0366 and in part by a National Science Foundation Fellowship.

† Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

$\alpha(G')=\theta(G')=r_1+r_2+1$. Hence we may assume no such c exists. But then we can partition C into two sets C_1 and C_2 such that $\alpha(G'(A\cup C_1))=r_1$, $\alpha(G'(B\cup C_2))=r_2$ (since if for every $c\in C_1$ $\alpha(G'(A\cup c))=r_1$ then $\alpha(G'(A\cup C_1))=r_1$ and similarly for B and C_2). Hence $\theta(G'(A\cup C_1))=r_1$, $\theta(G'(B\cup C_2)=r_2$ which implies $\theta(G')\leqq r_1+r_2$. Hence $\alpha(G')=\theta(G')=r_1+r_2$. This completes the proof of Lemma 1. □

LEMMA 2. *Let v be a point in G'. Let V be the vertex set of G'. Then $\alpha(G')=\theta(G')$ if and only if there exists a clique C in G' containing v such that $\alpha(G'(V-C))=\alpha(G')-1$.*

Proof. Let $\alpha(G')=\theta(G')=r$. Let $C_1, C_2, \cdots, C_r$ be a clique cover of G'. Let $v\in C_1$. Then $C_2, \cdots, C_r$ is a clique cover of $G'(V-C_1)$. Hence $\theta(G'(V-C_1))\leqq r-1$. Let $v_1, \cdots, v_r$ be a maximum set of independent vertices in G_1. Clearly, at most one v_i can be contained in C_1. Hence $\alpha(G'(V-C_1))\geqq r-1$. Since $\alpha(G'(V-C))=\theta(G'(V-C))$ we have $\alpha(G'(V-C))=r-1$. Hence we may let $C=C_1$.

Conversely let C be a clique in G' such that $\alpha(G'(V-C))=\alpha(G')-1$. Then since $\alpha(G'(V-C))=\theta(G'(V-C))$, $G'(V-C)$ has a clique cover containing $\alpha(G')-1$ cliques which implies G' has a clique cover containing $\alpha(G')$ cliques. Hence $\theta(G')=\alpha(G')$. □

LEMMA 3. *Let v_1 and v_2 be distinct vertices in G'. Suppose there does not exist a vertex v_3 distinct from v_1 and v_2 such that v_1 is adjacent to v_3 but v_2 is not adjacent to v_3. Then $\alpha(G')=\theta(G')$.*

Proof. Suppose v_1 is adjacent to v_2. Then any maximal clique in G' containing v_1 also contains v_2. Let V be the vertex set of G'. Then $\theta(G')=\theta(G'(V-v_2)=\alpha(G'(V-v_2))\leqq\alpha(G')\leqq\theta(G')$. Hence $\alpha(G')=\theta(G')$. Suppose next that v_1 is not adjacent to v_2. Let $r=\alpha(G'(V-v_2))$. By Lemma 2 there exists a clique C in $G'(V-v_2)$ containing v_1 such that $\alpha(G'(V-v_2-C))=r-1$. Suppose $\alpha(G'(V-C))=r-1$ also. Then $\theta(G'(V-C))=r-1$ so $r\geqq\theta(G')\geqq\alpha(G')\geqq\alpha(G'(V-v_2))=r$ which implies $\alpha(G')=\theta(G')$. Hence we assume $\alpha(G'(V-C))=r$. Then $G'(V-C)$ must contain a set of r independent vertices including v_2. We add v_1 to this set and obtain a set of $r+1$ independent vertices in G'. Now $\theta(G'(V-v_2))=r$ which implies $\theta(G')\leqq r+1$. Hence $r+1\leqq\alpha(G')\leqq\theta(G')\leqq r+1$ which implies $\alpha(G')=\theta(G')$. This completes the proof of Lemma 3. □

LEMMA 4. *Let v_1 and v_2 be distinct nonadjacent vertices in G'. Suppose any maximum sized independent set S of vertices in G' containing v_2 also contains v_1. Then $\alpha(G')=\theta(G')$.*

Proof. Let $G''=G'(V-v_2)$. By the induction hypothesis $\alpha(G'')=\theta(G'')$. Suppose $\alpha(G'')=\alpha(G')-1$. Clearly $\theta(G')\leqq\theta(G'')+1$. Hence $\alpha(G')\leqq\theta(G')\leqq\alpha(G')$ so $\alpha(G')=\theta(G')$ as desired. Hence we may assume $\alpha(G'')=\alpha(G')$. By Lemma 2 (applied to v_1 and G'') there exists a clique C in G'' containing v_1 such that $\alpha(G'(V-v_2-C))=\alpha(G'')-1=\alpha(G')-1$. We claim $\alpha(G'(V-C))=\alpha(G')-1$ also. For suppose not, then there exist a set of independent vertices in G' of size $\alpha(G')$ containing v_2 but no vertex in C. But since $v_1\in C$ and $\alpha(G')$ is the maximum size of a set of independent vertices in G' this is impossible. Now apply Lemma 2 to v_1, G' and C to obtain $\alpha(G')=\theta(G')$, as desired. □

We now continue with the proof of Theorem 1. Let v_1 be that rectangle in (the vertex set of) G' with lowest top row. If there are several such rectangles choose the one with lowest bottom row. Any remaining ties can be broken arbitrarily. Let L be the leftmost square in the top row of v_1 which is contained in another rectangle of G'. Let R be the rightmost square in the top row of v_1 which is contained in another rectangle of G'. Let I_1 be that portion of the top row of v_1 which lies between L and R inclusive. Note that any rectangle in G' intersecting v_1 must contain a square in

I_1. Let x be a lowest square not in P lying directly above a square in I_1. Let y be the square immediately below x. Suppose y lies in I_1. If y is contained in another rectangle v_2 of G' then because of the way we chose v_1, v_2 must contain the entire top row of v_1. Hence, if v_3 is any rectangle in G' which intersects v_1, v_3 intersects v_2 also. Therefore we are done by Lemma 3. Suppose y is contained in v_1 alone. Then removal of v_1 disconnects G since rectangles containing L are separated from rectangles containing R (we are using the fact that P is simply connected here). Hence in this case we are done by Lemma 1. Therefore we may assume y does not lie in I_1 which means I_1 does not consist of the entire top row of v_1 (else v_1 would not be a maximal rectangle) and we may assume without loss of generality that I_1 does not contain the upper left hand corner of v_1. Let I_2 be the row consisting of those squares on the same level as y and lying directly above squares in I_1. Let v_2 be a rectangle in G' intersecting v_1 but not I_2. Then v_2 must contain I_1 (see Fig. 1) and we may apply Lemma 3 as above.

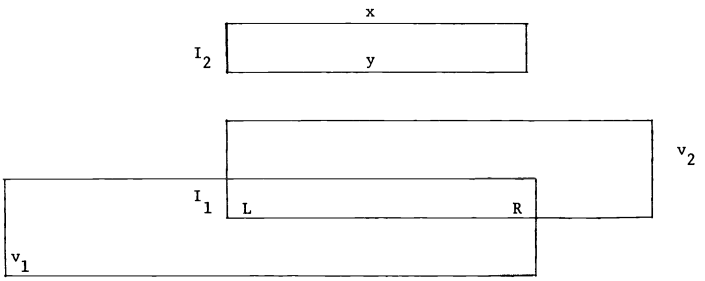


FIG. 1

Hence we may assume that all rectangles (except v_1) in G' intersecting I_1 intersect I_2 also. Let y' be that square in I_1 lying directly below y. Let I_3 consist of the column of squares lying between y' and y inclusive. Suppose a rectangle v_2 in G' intersects I_3 but not I_1 (see Fig. 2). Then all rectangles in G' intersecting v_1 (except v_1) intersect v_2 also (as otherwise they could not intersect I_2). Hence in this case we are done by Lemma 3. Let C be the clique in G' consisting of all rectangles in G' which contain y'. We may assume that the removal of C does not disconnect G' else we are done by Lemma 1. Hence we may assume all rectangles in G' intersecting I_1 but not containing y' lie on the same side of y', say the right (the proof of the other case proceeds analogously). Let v_2 be a rectangle containing L. Then v_2 must contain y'

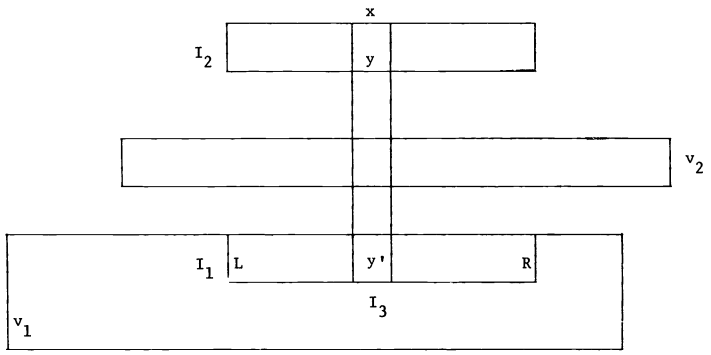


FIG. 2

also. Suppose any rectangle v_3 in G which intersects v_2 intersects v_1 also. Then we are done by Lemma 3 (with the roles of v_1 and v_2 interchanged). Hence we may let v_3 be a rectangle which intersects v_2 but not v_1. v_3 can not intersect I_3 as any rectangle intersecting I_3 also intersects v_1 (as we showed above). v_3 can not intersect v_2 to the left of I_3 as then removal of C disconnects G' (we are assuming G' contains a rectangle which intersects v_1 but not v_2 else we are done by Lemma 3). Hence v_3 must inersect v_2 to the right of I_3 as shown in Fig. 3. Note if v_4 is a rectangle which intersects v_1

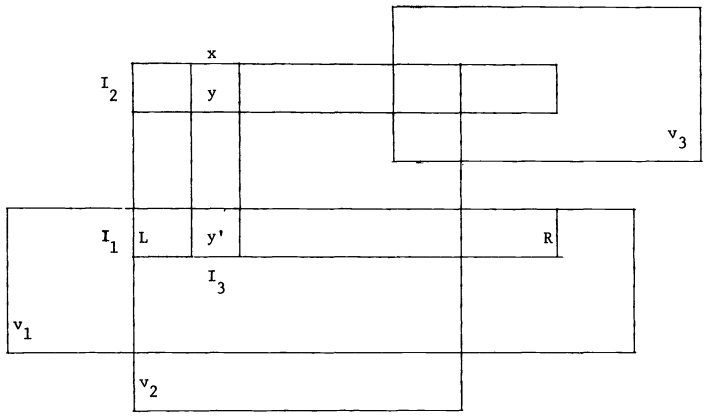


FIG. 3

but not v_3 then $v_4 \cap v_1 \subset v_2$. By Lemma 4 we may assume G' contains a maximum-sized independent set S of rectangles such that $v_3 \in S$, $v_1 \notin S$. Let $G'' = G'(V - v_2)$. Now $v_2 \notin S$ (as v_3 is) so S is independent in G'' also and $\alpha(G'') = \alpha(G')$. By the induction hypothesis $\alpha(G'') = \theta(G'')$. Apply Lemma 2 (with respect to v_1 and G'') to obtain a clique C' in G'' containing v_1 such that $\alpha(G'(V - v_2 - C')) = \alpha(G'') - 1$. We claim $v_2 \cup C'$ is a clique in G'. For let Z be a square contained in all the rectangles in C' (such a Z exists by Fact 1). Since $\alpha(G'(V - v_2 - C')) = \alpha(G'') - 1$ C' must contain a rectangle v_4 in S. Since C' is a clique and $v_1 \in C'$ v_4 must intersect v_1. Hence $v_4 \neq v_3$. Furthermore, v_4 does not intersect v_3 as $v_3 \in S$ (and S is an independent set). Hence as noted above $v_4 \cap v_1 \subset v_2$. Hence $Z \subset v_2$ which proves the claim. But then $\alpha(G'(V - v_2 \cup C')) = \alpha(G'') - 1 = \alpha(G') - 1$ so by Lemma 2 applied to v_1 and G' we have $\alpha(G') = \theta(G')$ which completes the proof of Theorem 1. □

Acknowledgments. The author wishes to thank Jeff Kahn, D. J. Kleitman and Dale Worley for their helpful comments.

REFERENCES

[1] C. BERGE, C. C. CHEN, V. CHVÁTAL AND C. S. SEOW, *Combinatorial properties of polyominoes*, preprint.

[2] V. CHVÁTAL, R. L. GRAHAM, A. F. PEROLD AND S. H. WHITESIDES, *Combinatorial designs related to the strong perfect graph conjecture*, Discrete Math., 26 (1979), pp. 83–92.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 3, September 1982

0196-5212/82/0303-0002 $01.00/0

A POINT-SYMMETRIC GRAPH THAT IS NOWHERE REVERSIBLE*

FRANK HARARY,† ANDREW VINCE‡ AND DALE WORLEY§

Abstract. It is the purpose of this note to investigate the relationships among four concepts relating to symmetry in graphs: point-symmetry, line-symmetry, arc-symmetry and reversibility; especially which of the first three properties do not imply reversibility. Holt has found a counterexample to one such question and we construct a counterexample to another using a Cayley graph. Both examples are nowhere reversible, a property which is stronger than nonreversibility.

1. Introduction. Let G be a connected graph. Its *automorphism group* $\Gamma(G)$ is defined as the group of line preserving permutations of the point set $V(G)$. Graph G is called *point-symmetric* if $\Gamma(G)$ is transitive on $V(G)$ and is called *line-symmetric* if $\Gamma(G)$ is transitive on its line set $E(G)$. If for the endpoints u, v of any line, there is an automorphism α such that $\alpha u = v$ and $\alpha v = u$, then G is called *reversible.* A slightly stronger notion is arc-symmetry. A graph G is *arc-symmetric* if, for any pair of (undirected) lines (u, v) and (u', v'), there is an automorphism α such that $\alpha u = u'$ and $\alpha v = v'$. (Arc-symmetry is called 1-transitivity in [2, p. 173].) We call a graph G *nowhere reversible* if there is no line (u, v) and automorphism α such that $\alpha u = v$ and $\alpha v = u$.

The following implications are immediate:

Fact 1. arc-symmetric $\Rightarrow$ reversible.

Fact 2. reversible $\Rightarrow$ point-symmetric.

Fact 3. arc-symmetric $\Rightarrow$ line-symmetric.

Fact 4. reversible and line-symmetric $\Leftrightarrow$ arc-symmetric.

The next two results are given in [2, p. 172] and [3].

Fact 5. If G is line-symmetric, but not bipartite, then G is point-symmetric.

Fact 6. If G is point- and line-symmetric with odd regularity degree, then G is arc-symmetric.

Two examples show that the converses of Facts 2 and 3 are false. The smallest graph that is line-symmetric but not point-symmetric (and thus not arc-symmetric) is shown in Fig. 1. The cubic graph in Fig. 2 is the smallest reversible graph that is not line-symmetric (and thus not arc-symmetric).

FIG. 1 *A line-symmetric but not point-symmetric graph.*

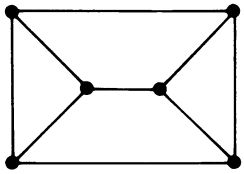


FIG. 2. *A reversible but not line-symmetric graph.*

* Received by the editors April 9, 1980, and in revised form October 5, 1981.

† Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48109.

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

§ Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

2. The counterexamples. The study of Facts 1 and 4 leads to the following less obvious questions:

Question 1. *What is the smallest graph that is both point- and line-symmetric, but nowhere reversible*?

Question 2. *What is the smallest point-symmetric graph that is nowhere reversible*?

Bouwer [1] found a large graph as an example for Question 1. More recently, Holt [4] has shown that the answer to Question 1 is at most 27 points. His 27-point graph H is constructed from the group

$$H = \langle x, y, x | x^9 = y^3 = z^2 = 1, y^{-1}xy = x^4, z^{-1} \times z = x^{-1}, yz = zy \rangle$$

and subgroup $S = \langle z \rangle$. The points of H are the 27 cosets of S. The (unordered) pair (S, Sx^7y) and all its images are the lines of H.

We next show that the answer to Question 2 is at most 21 points. The example is the Cayley graph G of the group B with generators T. The points of G are the elements of

$$B = \langle x, y | x^7 = y^3 = 1, xy = yx^2 \rangle.$$

Two points a and b are adjacent whenever $b = at$, where t is in $T = \{x, x^{-1}, y, y^{-1}, yx, y^{-1}x^3, y^{-1}x, yx^5\}$. Every Cayley graph is point-symmetric, but G is nowhere reversible.

THEOREM 1. *The graph G is nowhere reversible.*

Proof. Because H is a Cayley graph, any automorphism of G can be decomposed uniquely into a product of:

(a) an inner automorphism (one which is premultiplication by a fixed element of G), and

(b) an automorphism that fixes 1, which is either trivial or an outer automorphism (not an inner automorphism).

Clearly, some line is reversed by some inner automorphism iff T contains an element of order 2. Since T does not, any automorphism that reverses a line must have a nontrivial outer automorphism factor. We will demonstrate that G has no nontrivial automorphisms that fix 1, and so G is nowhere reversible.

Let β be an automorphism that fixes 1. Let G' be the link of 1, i.e., the subgraph of G induced by the vertices adjacent to 1. Figure 3 shows G'.

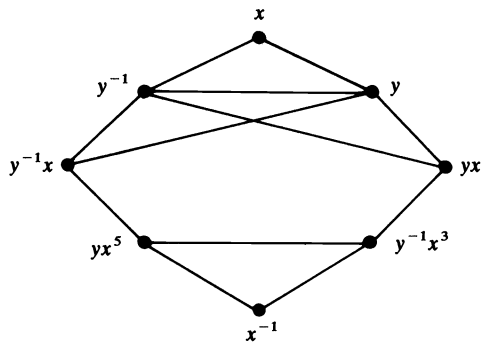


FIG. 3. *The link of the identity.*

Now, β, when restricted to G', must be an automorphism. Hence x and x^{-1} are also fixed points of β. Since G is point-symmetric, we have actually shown that if β fixes a, then β fixes ax. Thus, $x^0, x^1, \cdots, x^6$ must all be fixed by β. From G' we can

also deduce that β either fixes y and y^{-1} or it reverses them. But, y is adjacent to x^3 and y^{-1} is not, so β must fix y and y^{-1}. This shows that β fixes G.

We conclude by noting that the minimality conditions of both Questions 1 and 2 remain open. However, we conjecture that Holt's graph H and our graph G are indeed the smallest examples of Questions 1 and 2. Unfortunately, these questions will probably never be settled without an essentially exhaustive search for smaller examples. But, they may be solvable with reasonable effort when restricted to graphs generated by groups as H and G are.

Acknowledgment. We are grateful to Warren Brisley for kind advice and comments.

REFERENCES

[1] E. G. BOUWER, *Vertex and edge transitive, but not* 1*-transitive graphs*, Canad. Math. Bull., 13 (1970), pp. 231–237.

[2] F. HARARY, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.

[3] F. HARARY AND Z. MILLER, *On point-symmetric and arc-symmetric digraphs*, Nanta. Math., 10 (1977), pp. 50–52.

[4] D. HOLT, *A symmetric graph that is nowhere reversible.* J. Combin. Theory-B, to appear.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 3, September 1982

0196-5212/82/0303-0003 $01.00/0

THE RELATIONSHIP BETWEEN CONVEX GAMES AND MINIMUM COST SPANNING TREE GAMES: A CASE FOR PERMUTATIONALLY CONVEX GAMES*

DANIEL GRANOT† AND GUR HUBERMAN‡

Abstract. Notwithstanding the apparent differences between convex games and minimum cost spanning tree (m.c.s.t.) games, we show that there is a close relationship between these two types of games. This close relationship is realized with the introduction of the group of permutationally convex (p.c.) games. It is shown that a p.c. game has a nonempty core and that both convex games and m.c.s.t. games are permutationally convex.

Introduction. The core is the simplest and most intuitive solution concept for n-person cooperative games. It consists of all feasible payoff (or cost) vectors according to which no subset of the players can sever its cooperation with the rest of the players and be better off.

The complexity of the solution theory for cooperative games, as well as the deficiency of the core as a solution concept, is apparent when realizing that in general the core of a game may be empty. Nevertheless, being such a plausible solution concept, the core is considered as basic to the solution theory of cooperative games. In fact, other solution concepts gain support if it can be shown that they are in some way related to it. The core's existence (or its emptiness) is a very important property for any cooperative game. Its size and shape are crucial for any analysis, and it is usually the first thing one looks at when seeking a solution or when analyzing a cooperative game (see also [12]).

In light of the central role of the core in game theory, much effort has been devoted to characterize and study classes of games for which the core is not empty. Some examples of games with nonempty cores are convex games [11], [8], linear production games [10], market games [13] and the more recently introduced minimum cost spanning tree (m.c.s.t.) games [1], [3], [4]. A brief discussion of the relationship between convex and m.c.s.t. games is given below.

A convex game is defined as follows. Let $(N; c)$ be a (cost) cooperative game in characteristic function form[1], where $N = \{1, 2, \cdots, n\}$ is the set of players and $c: 2^N \to R$ is the characteristic function satisfying $c(\varnothing) = 0$. A cooperative game $(N; c)$ is *convex* (see [11]), if

$$c(S) + c(T) \geqq c(S \cup T) + c(S \cap T) \quad \text{for all } S, T \subseteq N. \tag{1}$$

The class of m.c.s.t. games is best introduced via the cablevision cost allocation problem (see [3], [4]). In the cablevision problem, the signals are initiated at a certain geographical point and are then transmitted through a tree network to various communities. Let us denote by c_{ij} the cost of transmitting signals between communities i, j, and let $C = (c_{ij})$ denote the symmetric cost matrix detailing the cost of connecting

* Received by the editors July 22, 1978 and in final revised form June 15, 1981.

† Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5.

‡ Graduate School of Business, University of Chicago, 1101 E. 58th Street, Chicago, Illinois 60637.

[1] In the game $(N; c)$ the characteristic function c is a cost function, unlike the usual case of a revenue characteristic function v, i.e., $c(S)$ is the cost incurred to the members of a coalition S by forming that coalition. Thus, we reverse the inequalities defining both various solution concepts (e.g., the core) and the properties of the characteristic functions (e.g., superadditivity and convexity).

any two communities. The cheapest transmission network is given by a minimum cost spanning tree associated with C.

The question of how to allocate the total cost of constructing the transmission network among the various communities can be answered by formulating this problem as a cooperative game, leading to the m.c.s.t. game formulation which is presented in § 2.

As it should be clear from the above discussion, convex games and m.c.s.t. games are basically different. This is obvious when one considers the size of the data required to specify a game in each class. A convex game is determined by specifying the $2^n - 1$ entries of the characteristic function, which is required to satisfy the convexity condition (1). On the other hand, an m.c.s.t. game is derived by merely specifying $(n+1)n/2$ entries in the symmetric cost matrix (c_{ij}).

In spite of the apparent difference between the two types of games, we demonstrate in this paper a close relationship between convex games and m.c.s.t. games. This relationship is illuminated here with the introduction of the class of permutationally convex (p.c.) games. A p.c. game is a generalization of a convex game which captures some of the properties of an m.c.s.t. game. Most importantly, the core of a p.c. game is not empty, and both convex games and m.c.s.t. games are permutationally convex. Finally, let us note that there are examples of convex m.c.s.t. games. These examples include Bird's minimal network game [1], the m.c.s.t. game that arises from existing minimal spanning tree networks (see Magiddo [9]) and Littlechild's airport game [7]. For further details see [5].

In the next section, we formally present m.c.s.t. games, motivate the introduction of p.c. games and prove the nonemptiness of their cores. In the last section, we show that m.c.s.t. games are indeed p.c.

2. Permutationally convex games: their definition and their cores. We first provide a brief review of m.c.s.t. games which will serve to motivate the introduction of permutationally convex (p.c.) games.

Let $N = \{1, \cdots, n\}$ denote the set of players, and let 0 designate a common supplier. The necessary data to define an m.c.s.t. game with n players are the entries of a cost matrix. Once the cost matrix $C = (c_{ij})$ $(i, j = 0, 1, 2, \cdots, n)$ is specified, the determination of the characteristic function of the corresponding m.c.s.t. game is as follows. For every set (coalition) $S \subseteq N$ construct a minimum cost graph, Γ_S, which spans $\{0\} \cup S$, and denote its cost by $c(S)$. (Of course, Γ_S is a tree, and hence the name of this group of games.) Given a characteristic function $c(\cdot)$ which maps all subsets of N into the real line, the core is the set of cost allocations (i.e., $x \in R^n$) which covers the total cost (i.e., $\sum_{i \in N} x_i = c(N)$) and which charges no subset (coalition) of N an amount higher than the cost of its independent operation (i.e., for all $S \subset N$, $\sum_{i \in S} x_i \leqq c(S)$).

The construction of an m.c.s.t. Γ_N induces a partial order $\succ$ on N, namely for $i, j \in N$ say that $i \succ j$ if node j is on the (unique) path connecting node i with the common supplier 0 in Γ_N. Throughout this paper we assume that the labeling of the nodes actually conforms to the partial order (i.e., if $i \succ j$, then $i > j$). Under the partial order $\succ$ each node $i \in N$ has one immediate predecessor $j(i) \in \{0, 1, \cdots, i-1\}$.

Suppose Γ_N was constructed using the following version of the greedy algorithm (see, e.g., [6]):

1. Let $\hat{N} = \{0\}$ and $\hat{E} = \varnothing$
2. If $\hat{N} = \{0\} \cup N$, stop. Else,
3. Choose a, b such that $a \in \hat{N}$, $b \in N \backslash \hat{N}$ and c_{ab} is minimal among

$$\{c_{xy} : x \in \hat{N},\ y \in N \backslash \hat{N}\}$$

4. Let $\hat{N} \leftarrow \hat{N} \cup \{b\}$, $\hat{E} \leftarrow \hat{E} \cup \{(a, b)\}$
5. Go to 2.

Upon termination the set $\hat{E} \subset \hat{N} \times \hat{N}$ is the edge set of Γ_N. The cost $c_{ij(i)}$ is the marginal cost of adding the node i to an existing m.c.s.t. of a subset of $\{0\} \cup N$. Denoting $MC_i \equiv c_{ij(i)}$ and $MC \equiv (MC_1, MC_2, \cdots, MC_n)$, we recall the main result on m.c.s.t. games, namely, that the vector MC is in the core of the corresponding m.c.s.t. game. reader is referred to [1], [2], [3], [4], [9] for further elaboration on m.c.s.t. games.

Convex games form another collection of games, which—like m.c.s.t. games—possess nonempty cores. Moreover, in both cases it is easy to compute at least one vertex of the core.

Denoting $[0] = \varnothing$ and $[k] = \{1, 2, \cdots, k\}$, let $x_i = c[i] - c[i-1]$. One of the main results on convex games is that if the game $(N; c)$ is convex then the vector x is a vertex of the core of $(N; c)$ for any ordering $1, 2, \cdots, n$ of the players.

To illuminate the analogy between the two types of games, recall that the characteristic function $c: 2^N \to R$ of a convex game satisfies

$$c(S \cup \{i\}) - c(S) \leqq c(T \cup \{i\}) - c(T) \tag{2}$$

for all $i \in N$, $T \subseteq S \subseteq N \backslash \{i\}$.

A reflection on (2) yields the following generalization. A game $(N; c)$ is *permutationally convex* if there exists a labeling of the players, say $1, 2, \cdots, n$, such that

$$c([k] \cup S) - c([k]) \leqq c([j] \cup S) - c([j]) \tag{3}$$

for all $S \subseteq N \backslash [k]$ and $k \geqq j$.

Any labeling $1, 2, \cdots, n$ of the players which satisfies (3) is a *permutationally convex order.* Note that a convex game is permutationally convex and that a permutationally convex game is convex if every labeling of the players is a permutationally convex order.

The nonemptiness of the core is our main result on permutationally convex games.

THEOREM. *Let $(N; c)$ be a permutationally convex game with a permutationally convex order $1, 2, \cdots, n$. Then the vector $x = (x_1, \cdots, x_n)$ is in the core of $(N; c)$, where $x_i = c[i] - c[i-1]$.*

Proof. It is immediate from the choice of $(x_1, \cdots, x_n)$ that $\sum_{i=1}^{n} x_i = c(N)$. To see that $\sum_{i \in S} x_i \leqq c(T)$ for all $T \subseteq N$, let $T = \{m_1, m_2, \cdots, m_r\}$, where $1 \leqq m_1 < m_2 < \cdots < m_r \leqq n$. Estimate $\sum_{i \in T} x_i$ using (3) as follows:

$$\sum_{i \in T} x_i = \sum_{j=1}^{r} (c([m_j]) - c([m_j - 1])) = c([m_r]) - c([m_r - 1]) + \sum_{j=1}^{r-1} (c([m_j] - c([m_j - 1])). \tag{4}$$

Apply (3) with $S = \{m_r\}$, $k = m_r - 1$ and $j = m_{r-1}$ to estimate

$$c([m_r]) - c([m_r - 1]) \leqq c(\{m_r\} \cup [m_{r-1}]) - c([m_{r-1}]),$$

which implies (using (4))

$$\sum_{i \in T} x_i \leqq c(\{m_r\} \cup [m_{r-1}]) - c([m_{r-1} - 1]) + \sum_{j=1}^{r-2} (c([m_j]) - c([m_j - 1])). \tag{5}$$

Apply (3) with $S = \{m_r\} \cup \{m_{r-1}\}$, $k = m_{r-1} - 1$ and $j = m_{r-2}$ to obtain (using (5))

$$\sum_{i \in T} x_i \leqq c(\{m_r\} \cup \{m_{r-1}\} \cup [m_{r-2}]) - c([m_{r-2} - 1]) + \sum_{j=1}^{r-3} (c([m_j]) - c([m_j - 1])).$$

A repeated application of similar estimates will result in

$$\sum_{i\in T} x_i \leqq c(\{m_r\}\cup\{m_{r-1}\}\cup\cdots\cup\{m_1\}) = c(T),$$

which completes the proof.

3. Minimum cost spanning tree games are permutationally convex. We prove that an m.c.s.t. game is permutationally convex, thereby providing another constructive core existence proof for this type of game. We recall that the players $1, 2, \cdots, n$ are labeled so that if j is on the (unique) path connecting i and 0 in the m.c.s.t. Γ_N, then $i > j$. Our objective is to show that any such labeling of the players is a permutationally convex order.

Define the cost matrices (c^i_{kl}) on the sets $(N\backslash[i])\cup\{0\}$ $(i = 1, \cdots, n-1)$ by

$$c^i_{kl} = \begin{cases} \min\limits_{j\in[i]} \{c_{k0}, c_{kj}\} & l = 0 \\ c_{kl} & l \neq 0 \end{cases} \Bigg\} k \in (N\backslash[i])\cup\{0\}. \tag{6}$$

The cost matrices (c^i_{kl}) determine the m.c.s.t. games $(N\backslash[i], c^i)$ $(i = 1, \cdots, n)$. From the definition of (c^i_{kl}) we have

$$c^k(S) \leqq c^j(S) \quad \text{for all } S \subseteq N\backslash[k] \text{ and } j \leqq k. \tag{7}$$

Next, apply the greedy algorithm to construct an m.c.s.t. for the graph obtained by collapsing $[j]\cup\{0\}$ to a single node and using the cost matrix (c^j_{kl}). Such an exercise results in the observation that

$$c[j] + c^j(S) = c([j]\cup S) \quad \text{for all } S \subseteq N\backslash[j]. \tag{8}$$

From (7) and (8) we have

$$c([k]\cup S) - c([k]) = c^k(S) \leqq c^j(S) = c([j]\cup S) - c([j]) \tag{9}$$

for all $S \subseteq N\backslash[k]$ and $j \leqq k$.

Now, (9) implies that $1, \cdots, n$ is a permutationally convex order and that m.c.s.t. games are indeed permutationally convex.

Acknowledgment. We are grateful to an anonymous referee for helpful suggestions which improved the presentation of this paper.

REFERENCES

[1] C. G. BIRD, *On cost allocation for a spanning tree: A game theoretic approach*, Networks, 6 (1976), pp. 335–350.

[2] A. CLAUS AND D. J. KLEITMAN, *Cost allocation for a spanning tree*, Networks, 3 (1973), pp. 289–304.

[3] D. GRANOT AND A. CLAUS, *Game theory application to cost allocation for a spanning tree*, Working Paper No. 402, Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, June 1976.

[4] D. GRANOT AND G. HUBERMAN, *On minimum cost spanning tree games*, Math. Prog., 21 (1981), pp. 1–18.

[5] ———, *The relationship between convex games and minimum cost spanning tree games: A case for permutationally convex games*, Discussion Paper 77-10-3, Simon Fraser University, Burnaby, B.C., June 1977.

[6] J. B. KRUSKAL, JR., *On the shortest spanning subtree of a graph and the traveling salesman problem*, Proc. Am. Math. Soc., 7 (1956), pp. 48–50.

[7] S. C. LITTLECHILD, *A simple expression for the nucleolus in a special case*, Internat. J. of Game Theory, 3 (1974), pp. 21–29.

[8] M. MASCHLER, B. PELEG AND L. S. SHAPLEY, *The kernel and bargaining set for convex games*, Internat. J. of Game Theory, 1 (1972), pp. 73–94.
[9] N. MEGIDDO, *Computational complexity and the game theory approach to cost allocation for a tree*, Math. Oper. Res., 3 (1978), pp. 189–196.
[10] G. OWEN, *On the core of linear production games*, Math. Prog., 9 (1975), pp. 356–370.
[11] L. S. SHAPLEY, *Cores of convex games*, Internat. J. of Game Theory, 1 (1971), pp. 11–26.
[12] L. S. SHAPLEY AND M. SHUBIK, *Game theory in economics—Chapter 6: characteristic function, core and stable set*, Report, R-904-NSF/6 (July 1973).
[13] ———, *On market games*, J. of Econ. Theory, 1 (1969), pp. 9–25.

SIAM J. ALG. DISC. METH
Vol. 3, No. 3, September 1982

0196-5212/82/0303-0004 $01.00/0

A SEMI-DEFINITE LYAPUNOV THEOREM AND THE CHARACTERIZATION OF TRIDIAGONAL D-STABLE MATRICES*

DAVID CARLSON†, B. N. DATTA‡ AND CHARLES R. JOHNSON§

Abstract. A new necessary and sufficient condition is given for an $n \times n$ complex matrix A to be stable. It involves a positive semi-definite image under a Lyapunov map and the real and imaginary parts of A. This condition is then used to characterize the real tridiagonal matrices which are D-stable, and those which are totally D-stable.

An $n \times n$ complex matrix A is said to be (positive) *stable* if each eigenvalue of A has positive real part. We shall be interested throughout in the *Lyapunov equation*

$$\tfrac{1}{2}(GA + A^*G) = H, \tag{1}$$

in which we usually assume

$$G \text{ is hermitian, positive definite} \tag{2}$$

and

$$H \text{ is (hermitian) positive semi-definite.} \tag{3}$$

The well-known [12] result of Lyapunov characterizes stable matrices in the following way.

THEOREM L. (i) *The matrix A is stable if and only if there exist positive definite matrices G and H satisfying* (1).

(ii) *Moreover, if hermitian G and H satisfy* (1), *then the positive definiteness of H implies that A is stable if and only if G is positive definite.*

While Lyapunov's theorem is, of course, of great importance in differential equations, the case of a semi-definite right-hand side in (1) seems also to arise often in applications (such as root location [5], [6]). In the event that (1) holds for given G satisfying (2) and H satisfying (3), A has no eigenvalues with negative real parts, but may have pure imaginary eigenvalues [3]. (The pure imaginary eigenvalues of A must have linear elementary divisors, viz., Jordan blocks of order 1 [3].) The controllability of (A^*, H) (cf. [9]) is then a necessary and sufficient condition for the stability of A [4], [16].

We shall give another characterization of the circumstances under which A is stable. While our condition will be seen to be equivalent to the controllability of (A^*, H), it will enable us to give simpler proofs than those in [4] and [16]. Moreover, in some other circumstances, it may be easier to work with than controllability. For example, in this paper, the condition will be applied to characterize the tridiagonal D-stable matrices, a goal which motivated the present work. The D-stable matrices arise [10], [11] in the stability analysis of general equilibrium economic systems with

* Received by the editors June 13, 1980, and in final and revised form October 15, 1981. This research was conducted at the Universidade Estadual de Campinas, and supported by grants 79/0085 and 79/0591 from the Fundação de Amparo à Pesquisa do Estado de São Paulo, Brazil.

† Mathematics Department, Oregon State University, Corvallis, Oregon 97331.

‡ Instituto de Matemática, Estatística, e Cienĉias da Computação, Universidade Estadual de Campinas, Campinas, Brazil. Current address: Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115.

§ Economics Department and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742. The work of this author was supported by the National Science Foundation under grant MCS 80-01611.

unknown (market specific) adjustment rates. They have not yet been fully characterized.

An important object which will provide the information necessary for a characterization is the skew-hermitian matrix S defined by

$$\tfrac{1}{2}(GA - A^*G) \equiv S. \tag{4}$$

From (1) and (4), it follows that

$$GA = H + S. \tag{5}$$

For a scalar λ to be an eigenvalue of A, there must exist a vector x such that

$$Ax = \lambda x, \qquad x \neq 0. \tag{6}$$

From (6), it follows that

$$GAx = \lambda Gx, \qquad x \neq 0, \tag{7}$$

which is equivalent to (6) if (2) holds. From (5) and (7) we obtain

$$(H+S)x = \lambda Gx, \tag{8}$$

$$x^*(H+S)x = \lambda \cdot x^*Gx \tag{9}$$

and

$$\mathrm{Re}\,(\lambda) \cdot x^*Gx = x^*Hx. \tag{10}$$

LEMMA 1. *Assuming* (1), (2), (3), (4), *and* (6), *we have that*

$$\mathrm{Re}\,(\lambda) = 0 \tag{11}$$

if and only if

$$\text{(a)}\quad Sx = \lambda Gx \quad \textit{and} \quad \text{(b)}\quad Hx = 0. \tag{12}$$

Proof. If (11) holds, it follows from (10) that

$$x^*Hx = 0, \tag{13}$$

which, because of (3), is equivalent to (12b). However, in view of (8), (12a) also follows, and we conclude that (11) implies (12). On the other hand, if (12) holds, then (11) follows from (10) since $x^*Gx > 0$ and $x^*Hx = 0$. This completes the proof. □

We note that under the hypotheses of Lemma 1, each pure imaginary eigenvalue of A is an eigenvalue of $G^{-1}S$ (whose eigenvalues are necessarily all pure imaginary) for which there is an eigenvector in the null space of H. To emphasize, A has a pure imaginary eigenvalue if and only if $G^{-1}S$ has an eigenvector in the null space of H, and, furthermore, the eigenvalues of $G^{-1}S$ (necessarily purely imaginary) whose eigenvectors lie in the null space of H are exactly the pure imaginary eigenvalues of A.

Remark. Let $G^{1/2}$ denote the positive definite square root of the positive definite matrix G, and $G^{-1/2}$ the inverse of $G^{1/2}$. The following are equivalent, as may be verified by simple algebraic manipulation:

(14) There is an $x \neq 0$ such that (12) holds.

(15) An eigenvector of $G^{-1}S$ lies in the null space of H.

(16) An eigenvector of SG^{-1} lies in the null space of HG^{-1}.

(17) An eigenvector of $G^{-1/2}SG^{-1/2}$ lies in the null space of $HG^{-1/2}$.

Our semi-definite generalization of Lyapunov's theorem is then

THEOREM 1. *Given matrix A, suppose that $H = \frac{1}{2}(GA + A^*G)$ for some positive definite matrix G and some positive semidefinite matrix H. Then A is stable if and only if*

(18) *No eigenvector of $G^{-1}S$ lies in the null space of H.*

Proof: Because of (2) and (3), it follows from (6) and (10) that all eigenvalues of A have nonnegative real parts. In view of Lemma 1, A is then stable if and only if there is no $x \neq 0$ satisfying (12). This is equivalent to the stated condition. □

Observe that if A is known to be nonsingular, we may replace (18) with

(18′) No eigenvector of $G^{-1}S$, corresponding to a nonzero eigenvalue, lies in the null space of H.

It is clear that if H is actually positive definite, then the condition of Theorem 1 is satisfied, since the null space of H consists only of the zero vector. Thus one direction of part (ii) of Theorem L is an immediate corollary of Theorem 1.

Let $\pi(A)$, $\nu(A)$, $\delta(A)$ denote (respectively) the number of eigenvalues of A, counting algebraic multiplicity, with positive, negative, and zero real parts. We have

COROLLARY 1. *Assuming* (1), (2), *and* (3), *we have*

$$\pi(A) \geqq \operatorname{rank}(H) \qquad (= \pi(H)),$$

with equality holding if and only if there is a basis of the null space of H consisting of eigenvectors of $G^{-1}S$.

Proof: In view of (2) and the fact that $\nu(A) = 0$, the inequality of the corollary is equivalent to

$$\delta(A) \leqq \delta(H).$$

As pure imaginary eigenvalues of A have linear elementary divisors, this now follows from Lemma 1. □

The inequality of Corollary 1 may also be found in [3]. Since, for a nonsingular real matrix A, $\delta(A)$ cannot be odd, we further conclude:

COROLLARY 2. *If a nonsingular real matrix A satisfies* (1), *while G satisfies* (2) *and H satisfies* (3), *then, if* $\operatorname{rank}(H) \geq n-1$, *it follows that A is stable.*

To show the equivalence of our condition with (A^*, H) controllable, we note that Hautus [9] has proved that (A^*, H) is controllable if and only if

$$Ax = \lambda x, \qquad H^*x = 0 \Rightarrow x = 0.$$

In our case, $A = G^{-1}(H+S)$, H is hermitian, and clearly

$$Ax = \lambda x, \qquad H^*x = 0 \Leftrightarrow G^{-1}Sx = \lambda x, \qquad Hx = 0.$$

We have proved:

LEMMA 2. *Assuming $A = G^{-1}(H+S)$ with H hermitian, (A^*, H) is controllable if and only if* (18) *holds.*

The equation (1) has also been studied for hermitian G and H, with no restriction that G satisfy (2). In this case, if H is positive definite, then G is nonsingular and we have "equality of inertias" for A and G:

$$\pi(A) = \pi(G), \qquad \nu(A) = \nu(G), \qquad \delta(A) = \delta(G) = 0 \tag{19}$$

([14], [15]). If we assume just that H satisfies (3), then it is known [4], [16], [2, example, p. 240] that the controllability of (A^*, H) is sufficient but not necessary for

G to be nonsingular and the inertias of A and G to be equal, i.e., (19). Clearly we have:

COROLLARY 3. *Suppose that* $H=\frac{1}{2}(GA+A^*G)$ *holds for nonsingular hermitian* G *and positive semidefinite* H. *If also* (18) *holds, then the inertias of* A *and* G *are equal.*

We next turn to the consideration of D-stable matrices which have received considerable study [10], [11] but have not been characterized. We will henceforth assume all matrices to be real. A stable matrix A is called D-stable if DA is stable for all positive diagonal matrices. Such a matrix is further called *totally D-stable* if each of its principal submatrices is D-stable. A well-known necessary condition for D-stability is that all principal minors be nonnegative and at least one of each size be positive [10]. We call the class of all such matrices P_0^+. That all principal minors be positive is necessary for total D-stability, and we call this class P^+. Neither condition is, in general, sufficient, and one sufficient condition for D-stability (in fact for total D-stability) is that there exist a diagonal solution G to (1) with H positive definite. In case $A=(a_{ij})$ is *tridiagonal* ($a_{ij}=0$ whenever $|i-j|>1$), we characterize total D-stability, because of the happy coincidence of necessary and sufficient conditions, and also characterize D-stability.

We first consider totally D-stable tridiagonal matrices.

THEOREM 2. *For a tridiagonal matrix* A, *the following conditions are equivalent*:

(i) *There is a positive diagonal matrix* D *such that* $DA+A^TD$ *is positive definite*;

(ii) A *is totally D-stable*; *and*

(iii) $A\in P^+$.

Proof. The implications (i)$\Rightarrow$(ii) and (ii)$\Rightarrow$(iii) are known (and straightforward) for general matrices. That (iii)$\Rightarrow$(i) holds for tridiagonal matrices follows from a construction of D. First, assume A is irreducible (cf. [8] or [13]) so that $a_{i,i+1}a_{i+1,i}\neq 0$ for $i=1,\cdots,n-1$. Then, define $D=\operatorname{diag}(d_1,\cdots,d_n)$ sequentially by

$$d_1=1, \qquad d_{i+1}=d_i\frac{|a_{i,i+1}|}{|a_{i+1,i}|}, \quad i=1,\cdots,n-1. \tag{20}$$

Now, in DA, the absolute values of the $(i,i+1)$, and $(i+1,i)$ entries are equated whence it follows that in $\frac{1}{2}(DA+A^TD)$, the $(i,i+1)$ and $(i+1,i)$ entries either agree with those of DA or are both zero. Since $A\in P^+$, DA is also, and it is then easy to check that $DA+A^TD\in P^+$ since each irreducible direct summand of $DA+A^TD$ agrees with the corresponding submatrix of DA. But this means that, since it is symmetric, $DA+A^TD$ is also positive definite.

To prove the theorem when A is reducible, assume first that for some k, $1\leqq k\leqq n-1$,

$$a_{k+1,k}a_{k,k+1}=0, \quad \text{but} \quad a_{i+1,i}a_{i,i+1}\neq 0, \quad i=1,\cdots,n-1, \quad i\neq k. \tag{21}$$

Let A_{11} (A_{22}) denote the principal submatrix contained in the first k (last $n-k$) rows and columns of A. By (21), for $i=1,2$, A_{ii} is irreducible, so that there exists a positive diagonal matrix D_i so that $D_iA_{ii}+A_{ii}^TD_i$ is positive definite. If $a_{k+1,k}=a_{k,k+1}=0$, clearly $D=D_1\oplus D_2$ is a positive diagonal matrix for which $DA+A^TD$ is positive definite. If $a_{k,k+1}\neq 0$, define $D_\varepsilon=D_1\oplus\varepsilon D_2$, $\varepsilon>0$. The matrix $H_\varepsilon=D_\varepsilon A+A^TD_\varepsilon$ is positive definite for all sufficiently large $\varepsilon>0$. To see this, observe that H_ε has the form

$$H_\varepsilon=\begin{pmatrix} H_{11} & H_{12} \\ H_{12}^T & \varepsilon H_{22}\end{pmatrix}$$

with H_{11}, H_{22} positive definite. Now H_ε is positive definite if and only if H_{11} and the Schur complement $\varepsilon H_{22}-H_{12}^TH_{11}^{-1}H_{12}$ are both positive definite; and this Schur

complement is positive definite for sufficiently large $\varepsilon > 0$. Similarly, if $a_{k+1,k} \neq 0$, define $D_\varepsilon = \varepsilon D_1 \oplus D_2$, $\varepsilon > 0$.

An inductive argument based on the above completes the proof for arbitrary reducible A. □

In order to characterize the tridiagonal matrices which are D-stable (and not necessarily totally D-stable), it is clearly sufficient to characterize the irreducible tridiagonal matrices which are D-stable. We will consider three cases: these deal with the irreducible tridiagonal matrices $A = (a_{ij})$ for which

$$a_{i,i+1}a_{i+1,i} > 0, \qquad i = 1, 2, \cdots, n-1, \tag{22}$$

$$a_{i,i+1}a_{i+1,i} < 0, \qquad i = 1, 2, \cdots, n-1, \tag{23}$$

and the general case, when neither (22) nor (23) need hold. We shall define an $n \times n$ tridiagonal real matrix A to be *skew* if $a_{ii} = 0$, $i = 1, \cdots, n$, and (23) holds. Before giving our characterization of the irreducible tridiagonal D-stable matrices, we need two lemmas, one dealing with skew and the other dealing with skew-symmetric matrices.

LEMMA 3. *Suppose that $n \times n$ tridiagonal matrices A and B are skew and irreducible. Then there exist a positive diagonal matrix E and a nonsingular diagonal matrix F for which $B = FEAF^{-1}$.*

Proof. Given tridiagonal matrices $A = (a_{ij})$ and $B = (b_{ij})$, skew and irreducible, it is sufficient to exhibit nonsingular diagonal matrices X and Y for which $B = XAY$ and XY is positive diagonal, for then $F = Y^{-1}$ and $E = XY$ satisfy the conditions of the Lemma.

Let $X = \operatorname{diag}(x_1, \cdots, x_n)$ and $Y = \operatorname{diag}(y_1, \cdots, y_n)$. We have $B = XAY$ if and only if

$$\text{and} \qquad \begin{aligned} a_{i,i+1}x_i y_{i+1} &= b_{i,i+1}, \\ a_{i+1,i}x_{i+1}y_i &= b_{i+1,i}, \end{aligned} \qquad i = 1, 2, \cdots, n-1.$$

But if we now choose $x_1 > 0$, we obtain sequentially

$$\text{and} \qquad \begin{aligned} y_{2k} &= \frac{b_{2k-1,2k}}{a_{2k-1,2k}x_{2k-1}}, \\ x_{2k+1} &= \frac{b_{2k-1,2k}^2}{a_{2k+1,2k}y_{2k}}, \end{aligned} \qquad k = 1, 2, \cdots,$$

and if we choose $y_1 > 0$, we obtain sequentially,

$$\text{and} \qquad \begin{aligned} x_{2k} &= \frac{b_{2k,2k-1}}{a_{2k,2k-1}y_{2k-1}}, \\ y_{2k+1} &= \frac{b_{2k,2k+1}}{a_{2k,2k+1}x_{2k}}, \end{aligned} \qquad k = 1, 2, \cdots.$$

For X and Y determined in this way, we have $B = XAY$ and XY positive diagonal. □

Given positive integers n and p, $1 \leqq p \leqq n$, following [13, p. 9] we define

$$Q_{n,p} = \{\omega = (i_1, \cdots, i_p) | 1 \leqq i_1 < \cdots < i_p \leqq n\}.$$

For vector x, we define $\omega(x) = (i_1, \cdots, i_p) \in Q_{n,p}$ to be the sequence of positions of zero components of the vector. (If no components of x are zero, we define $\omega(x) = \varnothing$.)

LEMMA 4. (i) *Suppose tridiagonal matrix A is skew-symmetric and irreducible, and that $Ax = \lambda x$ for some $\lambda \neq 0$ and $x \neq 0$. If $\omega(x) = (i_1, \cdots, i_p) \neq \varnothing$, then*

$$(24) \qquad i_1 \geqq 3, \quad i_2 - i_1 \geqq 3, \quad \cdots, \quad i_p - i_{p-1} \geqq 3, \quad n - i_p \geqq 2.$$

(ii) *Suppose either $\omega = \varnothing$ or $\omega \in Q_{n,p}$ satisfying* (24), *and suppose $\lambda \neq 0$ is imaginary. Then there exist a tridiagonal matrix A, skew-symmetric and irreducible, and an $x \neq 0$ with $\omega(x) = \omega$, for which $Ax = \lambda x$.*

Proof. We may assume $n \geqq 2$. Let tridiagonal A be skew-symmetric and irreducible. For convenience of notation, we let $a_{i,i+1} = b_i$, $a_{i+1,i} = -b_i$, $i = 1 \cdots, n-1$. Now $Ax = \lambda x$ is equivalent to

$$(e_1) \qquad 0 = \lambda x_1 - b_1 x_2,$$

$$(e_i) \qquad 0 = b_{i-1} x_{i-1} + \lambda x_i - b_i x_{i+1}, \quad i = 2, \cdots, n-1,$$

$$\cdots$$

$$(e_n) \qquad 0 = b_{n-1} x_{n-1} + \lambda x_n.$$

Proof of (i). Suppose $Ax = \lambda x$, $\lambda \neq 0$, but $\omega(x) \neq \varnothing$. As A is tridiagonal and irreducible, it follows from (e_1)–(e_n) that if $x_i = x_{i+1} = 0$ for any $i = 1, \cdots, n-1$, then $x = 0$. If $x_1 = 0$ or $x_2 = 0$, then by (e_1), $x_1 = x_2 = 0$, and $x = 0$, i.e., if $x \neq 0$, $i_1 \geqq 3$. If $x_{n-1} = 0$ or $x_n = 0$, then by (e_n), $x_{n-1} = x_n = 0$, and $x = 0$, i.e., if $x \neq 0$, $n - i_p \geqq 2$. Finally, if $x \neq 0$ and $x_i = 0$, then $x_{i+1} \neq 0$ and $x_{i+2} = \lambda x_{i+1}/b_{i+1} \neq 0$, i.e., $i_k - i_{k-1} \geqq 3$.

Remark. The argument for (i) clearly can be used to show that for any irreducible tridiagonal A, with all nonzero diagonal entries, and any x with $\omega(x) \neq \varnothing$ for which $Ax = 0$, (24) must hold.

Proof of (ii). The proof is based on the following observation. For imaginary λ, μ with $\lambda\mu < 0$, and nonzero complex u, if real number $b \neq 0$ satisfies $b^2 + \lambda\mu = 0$, and $v = \mu u/b$, then $\mu u - bv = 0$, $bu + \lambda v = 0$. If real $b \neq 0$ satisfies $b^2 + \lambda\mu < 0$, and $v = u/b$, then $\mu u - bv = 0$, $(bu + \lambda v)\lambda/u = (b^2 + \mu\lambda)\lambda/\mu < 0$, and $(bu + \lambda v)/v$ is imaginary. Note that if u is real, v is pure imaginary, and vice versa.

We first prove (ii) for $\omega = \varnothing$. For $n = 2$, we take $\mu = \lambda$; then $\lambda\mu = \lambda^2 < 0$. For any nonzero real b_1 satisfying $b_1^2 + \lambda^2 = 0$, any nonzero complex $x_1 = u$, and nonzero $x_2 = v = \lambda x_1/b_1$, we have a solution $x = (x_1, x_2)^T$ by our observation. For $n = 3$, we take $\mu_1 = \lambda$, any nonzero real b_1 satisfying $b_1^2 + \lambda^2 < 0$, any nonzero complex $x_1 = u$, and nonzero $x_2 = v = \lambda x_1/b_1$. Now (x_1, x_2) is a solution of (e_1), and $\mu_2 = (b_1 x_1 + \lambda x_2)/x_2$ is imaginary, satisfying $\mu_2\lambda < 0$. We may rewrite (e_2) as

$$0 = \frac{(b_1 x_1 + \lambda x_2)}{x_2} \cdot x_2 - b_2 x_3.$$

Now for any nonzero real b_2 satisfying $b_2^2 + \lambda\mu_2 < 0$, $x_2 = u$, and nonzero $x_3 = v = \mu_2 x_2/b_2$, we have $x = (x_1, x_2, x_3)^T$ as a solution also of (e_2) and (e_3). An inductive proof for arbitrary n along these lines is easy, and will be omitted. The proof of (ii) for $\omega = \varnothing$ (and $n \geqq 2$) is complete. Observe that x may be chosen so that $x_1, x_2, \cdots, x_n$ alternate between real and pure imaginary.

We must yet prove (ii) for $\omega \in Q_{n,p}$ satisfying (24). We shall give a proof for $p = 1$. A similar inductive argument for arbitrary p is easy, and will be omitted. Let $\omega = (i)$. From (24), we must have $i \geqq 3$ and $n - i \geqq 2$. We shall consider matrices and vectors

of the forms

$$A^{(1)}=\begin{pmatrix} 0 & b_1 & & & 0 \\ -b_1 & 0 & \ddots & & \\ & \ddots & \ddots & 0 & b_{i-2} \\ & & \ddots & -b_{i-2} & 0 \\ 0 & & & & \end{pmatrix}, \qquad A^{(2)}=\begin{pmatrix} 0 & b_{i+1} & & & 0 \\ -b_{i+1} & 0 & \ddots & & \\ & \ddots & \ddots & 0 & b_{n-1} \\ & & \ddots & -b_{n-1} & 0 \\ 0 & & & & \end{pmatrix},$$

$$x^{(1)}=\begin{pmatrix} x_1 \\ \vdots \\ x_{i-1} \end{pmatrix}, \qquad x^{(2)}=\begin{pmatrix} x_{i+1} \\ \vdots \\ x_n \end{pmatrix}.$$

Now $A^{(1)}$, $A^{(2)}$, $x^{(1)}$, $x^{(2)}$ each have at least two rows. By what we have just proved for $\omega=\varnothing$, for $k=1,2$ there exist $A^{(k)}$, $x^{(k)}$ for which $A^{(k)}x^{(k)}=\lambda x^{(k)}$, $\omega(x^{(k)})=\varnothing$, and for which the components of $x^{(k)}$ alternate between real and pure imaginary. This is sufficient to show that for any b_{i-1}, b_i, and $x_i=0$, and

$$A=\left(\begin{array}{ccc|c|cc} & & & 0 & & \\ & A^{(1)} & & \vdots & & 0 \\ & & & 0 & & \\ & & & b_{i-1} & & \\ \hline 0\cdots 0 & & -b_{i-1} & 0 & b_i & 0\cdots 0 \\ \hline & & & -b_i & & \\ & 0 & & 0 & & A^{(2)} \\ & & & \vdots & & \\ & & & 0 & & \end{array}\right), \qquad x=\begin{pmatrix} x^{(1)} \\ \hline 0 \\ \hline x^{(2)} \end{pmatrix}$$

$(e_1), \cdots, (e_{i-1}), (e_{i+1}), \cdots, (e_n)$ hold. Now (e_i) becomes

(e_i') $\qquad 0=b_{i-1}x_{i-1}-b_ix_{i+1}.$

If both x_{i-1} and x_i are real, or both are imaginary, we may choose nonzero reals b_{i-1}, b_i so that (e_i') holds. If not, we merely replace $x^{(2)}$ by $ix^{(2)}$, and then we may choose b_{i-1}, b_i real as required. We have A real, tridiagonal, irreducible and skew-symmetric; $x\neq 0$, with $\omega(x)=(i)$, and $Ax=\lambda x$. □

In the characterization of irreducible tridiagonal matrices which are D-stable, the paragraph which follows gives a common outline for the proofs of D-stability in all three cases (i.e., matrices satisfying (22) and (23), and the general case).

Given $A\in P_0^+$, tridiagonal and irreducible, and any positive diagonal matrix E, there exists a positive diagonal matrix F such that $F(EA)F^{-1}=B$ satisfies $|b_{i,i+1}|=|b_{i+1,i}|$, $i=1,\cdots,n-1$. (To obtain F, let F^2 be defined by (20) applied to EA; then $F^{-1}(F^2EA)F^{-1}=FEAF^{-1}=B$.) We have

$$\tfrac{1}{2}(IB+B^TI)=H, \qquad \tfrac{1}{2}(IB-B^TI)=S,$$

with H symmetric and S skew-symmetric. As $B\in P_0^+$, and the irreducible principal submatrices of H are just principal submatrices of B, H is positive semidefinite, i.e., H satisfies (3). Of course $G=I$ satisfies (2). If, for each positive diagonal E, (18′) holds, and EA is stable, then A is D-stable.

LEMMA 5. *Let A be an irreducible tridiagonal matrix satisfying* (22). *Then A is D-stable (and in fact totally D-stable) if and only if $A\in P_0^+$.*

Proof. We have already noted that $A \in P_0^+$ is a necessary condition for D-stability. Suppose that $A \in P_0^+$, and that E and F are positive diagonal matrices as indicated above. Now $B = FEAF^{-1}$ is symmetric, so that $B = H$ and $S = 0$. As S has no nonzero eigenvalues, (18′) holds, and B and EA are stable. As this is true for every positive diagonal E, A is D-stable.

We give a second proof, showing that A is totally D-stable. By (20), we can choose a positive diagonal matrix D so that $DA + A^TD$. Then $\frac{1}{2}(DA + A^TD) = DA \in P_0^+$ since A is. But a symmetric matrix in P_0^+ is positive definite; it is positive semidefinite by definition, and has positive determinant. Now A is totally D-stable by Theorem 2. □

For each $\omega = (i_1, \cdots, i_p) \in Q_{n,p}$, either (24) or its negation,

$$(25) \qquad \begin{aligned} &i_1 < 3, \quad \text{or} \quad i_{h+1} - i_h < 3 \quad \text{for some } h = 1, 2, \cdots, p-1, \\ &\text{or} \quad i_p > n-2, \end{aligned}$$

holds. For matrix A, let $\phi(A) = (i_1, \cdots, i_p)$ be the sequence of indices of diagonal entries which are *not* zero. (If all diagonal entries of A are zero, we define $\phi(A) = \varnothing$. If $A \in P_0^+$, clearly $\phi(A) \neq \varnothing$.)

LEMMA 6. *Let $A \in P_+^0$ be irreducible and tridiagonal, satisfying* (23). *Then A is D-stable if and only if* (25) *holds for $\phi(A)$.*

Proof. Suppose $A \in P_0^+$, and that (25) holds for $\phi(A) = (i_1, \cdots, i_p)$. Let E and F be positive diagonal matrices as indicated above. In this case, H is diagonal, and S is irreducible. Let $Sx = \lambda x$, $\lambda \neq 0$.

By Lemma 4(i), either $\omega(x) = \varnothing$ or $\omega(x) = (j_1, \cdots, j_q) \neq \varnothing$, satisfying (24). As $\phi(A) = \phi(H) \neq \varnothing$, $Hx \neq 0$ if $\omega(x) = \varnothing$ also. Suppose $\omega(x) = (j_1, \cdots, j_q) \neq \varnothing$ satisfying (24). Recall that we are assuming that $\phi(A) = (i_1, \cdots, i_p)$ satisfies (25). If $i_1 < 3$, either $a_{11} \neq 0$ or $a_{22} \neq 0$, and as $j_1 \geq 3$, $Hx \neq 0$; similarly if $i_p > n-2$. If $i_{k+1} - i_k \leqq 2$ for some $k = 1, \cdots, p-1$, either $i_{k+1} = i_k + 1$ or $i_{k+1} = i_k + 2$, and either $a_{i_k i_k} a_{i_k+1, i_k+1} \neq 0$ or $a_{i_k i_k} a_{i_k+2, i_k+2} \neq 0$. As $\omega(x)$ satisfies (24), the zero components of x occur at least 3 positions apart. We must have some $a_{ii} x_i \neq 0$, and hence $Hx \neq 0$. In every possible situation, $Hx \neq 0$; thus by Lemma 1, $B = F(EA)F^{-1}$ is stable, hence so is EA. As this holds for every positive diagonal matrix E, A is D-stable.

Suppose now that $A \in P_0^+$, and with (24) instead of (25) holding for $\phi(A) = (i_1, \cdots i_p)$.

By Lemma 4(ii), there exist an imaginary $\lambda \neq 0$, a real tridiagonal B, skew-symmetric and irreducible, and an $x \neq 0$ with $\omega(x) = \phi(A)$, for which $Bx = \lambda x$.

Now $D = \operatorname{diag}(a_{11}, \cdots, a_{nn})$ is positive semi-definite, and $A - D$ is tridiagonal, skew and irreducible. By Lemma 3, there exist positive diagonal E and nonsingular diagonal F for which $B = FE(A-D)F^{-1}$. We have $FEAF^{-1} = ED + B$, and, since $B^T = -B$,

$$\tfrac{1}{2}(I(ED+B) + (ED+B)^T I) = ED, \qquad \tfrac{1}{2}(I(ED+B) - (ED+B)^T I) = B,$$

with ED satisfying (3). As $\omega(x) = \phi(A) = \phi(ED)$, $EDx = 0$, while $Bx = \lambda x$. By Theorem 1, $FEAF^{-1}$ and thus EA are not stable, so that A is not D-stable. □

Before stating and proving our characterization in the general case, we must define and examine a decomposition of any irreducible tridiagonal A related to the decomposition of $B = FEAF^{-1}$ already discussed. We write $A = H + S$, where H and

S are also tridiagonal, and

$$h_{ij} = \begin{cases} a_{ij} & \text{if } i = j, \quad \text{or } i \neq j \quad \text{and } a_{ij}a_{ji} > 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$s_{ij} = \begin{cases} a_{ij} & \text{if } i \neq j \quad \text{and } a_{ij}a_{ji} < 0, \\ 0 & \text{otherwise.} \end{cases}$$

We may partition S and H as

$$S = \begin{pmatrix} S_1 & & 0 \\ & \ddots & \\ 0 & & S_k \end{pmatrix}, \tag{26}$$

where each S_j is either irreducible or 1×1 and zero, and, conformably,

$$H = \begin{pmatrix} H_{11} & H_{12} & & 0 & 0 \\ H_{21} & H_{22} & \ddots & 0 & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & 0 & \ddots & H_{k-1,k-1} & H_{k-1,k} \\ 0 & 0 & & H_{k,k-1} & H_{k,k} \end{pmatrix} \tag{27}$$

where

$$H_{jj} = \begin{pmatrix} h_1^{(j)} & & 0 \\ & \ddots & \\ 0 & & h_{n_j}^{(j)} \end{pmatrix}, \quad H_{j,j+1} = \begin{pmatrix} 0 & 0 \\ & \\ b_j & 0 \end{pmatrix}, \quad H_{j+1,j} = \begin{pmatrix} 0 & c_j \\ & \\ 0 & 0 \end{pmatrix}, \tag{28}$$

and $b_j c_j > 0$. We shall also assume A partitioned conformably with S. We note that if (22) holds, $A = H$, $S = 0$, and $k = n$; if (23) holds, H is diagonal and S is irreducible, i.e., $k = 1$; and if neither (22) nor (23) holds, then at least one $S_j \neq 0$, and $1 < k < n$.

If $1 < k < n$, we shall be interested in the $\phi(A_j)$ (defined analogously for each $j = 1, \cdots, k$) for those j for which $S_j \neq 0$. If $S_j \neq 0$ and $S_{j+1} \neq 0$, we shall call the last diagonal entry of A_j and the first diagonal entry of A_{j+1} *transition entries*; diagonal entries which are not transition entries shall be called *interior entries*. We will say that $\phi(A_j)$ *satisfies* (25) *for interior entries* if it satisfies (25) *except that* transition entries cannot be used to satisfy (25). Thus, if the first diagonal entry of A_j is a transition entry, then $i_1 < 3$ is replaced in (25) by $2 \in \phi(A_j)$, and if the last diagonal entry of A_j is a transition entry, $i_p > n-2$ is replaced by $n-1 \in \phi(A_j)$. We will say that $\phi(A)$ *satisfies* (25) *for interior entries* if at least one $\phi(A_j)$ satisfies (25) for interior entries. We shall call $a_{i,i}a_{i+1,i+1} - a_{i,i+1}a_{i+1,i}$ a *transition minor* if a_{ii} and $a_{i+1,i+1}$ are transition entries (and $a_{i,i+1}a_{i+1,i}, i > 0$).

If D_1, D_2 are nonsingular diagonal matrices, decomposition of D_1AD_2 will yield the same partitioning as that of A, and $\phi((D_1AD_2)_j) = \phi(A_j), j = 1, \cdots, k$. A transition minor of D_1AD_2 will be zero if and only if the corresponding transition minor of A is zero.

THEOREM 3. *Let $A \in P_0^+$ be irreducible and tridiagonal. Then A is D-stable if and only if*

(29) *$\phi(A)$ satisfies* (25) *for interior entries, or*

(30) *at least one transition minor is nonzero, or*

(31) $S_1 = 0$ *or* $S_k = 0$ *or at least two successive* $S_j = 0$.

Proof. Let $A \in P_0^+$ be irreducible and tridiagonal. Let E be a positive diagonal matrix, and let F and $B = FEAF^{-1} = H + S$ as before. Partition H and S as in (26)–(28). Suppose $Sx = \lambda x$ and $Hx = 0$ for some $\lambda \neq 0$ and for $x = (x_j)$, partitioned conformably with S. For each j, we have $S_j x_j = \lambda x_j$; and either $x_j = 0$ or x_j is an eigenvector of S_j associated with the eigenvalue λ. As $\lambda \neq 0$, if $S_j = 0$, also $x_j = 0$. If $x_j \neq 0$, we know that $\omega(x_j)$ satisfies (24). As $Hx = 0$, we have, analogous to (e_1)–(e_n),

(f_1) $\quad 0 = H_{11}x_1 + H_{12}x_2,$

(f_j) $\quad 0 = H_{j,j-1}x_{j-1} + H_{jj}x_j + H_{j,j+1}x_{j+1}, \qquad j = 2, \cdots, k-1,$

(f_k) $\quad 0 = H_{k,k-1}x_{k-1} + H_{k,k}x_k.$

Whenever $x_j \neq 0$, $\omega(x_j)$ satisfies (24), and $H_{j-1,j}$ and $H_{j+1,j}$ have the form given in (28), thus $H_{j-1,j}x_j \neq 0$ and $H_{j+1,j}x_j \neq 0$. It follows from (f_1)–(f_k) that whenever $x_j = x_{j+1} = 0$, $x = 0$; and if $x_1 = 0$ or $x_k = 0$, $x = 0$. If (31) holds, then $x = 0$.

Let $x_j = (x_1^{(j)}, \cdots, x_{n_j}^{(j)})^T$, $j = 1, \cdots, k$; assuming $S_j \neq 0$, $n_j > 1$, and equation (f_j) is equivalent to the system

(g_1^j) $\quad c_{j-1}x_{n_{j-1}}^{(j-1)} + h_1^{(j)}x_1^{(j)} = 0,$

(g_l^j) $\quad h_l^{(j)}x_l^{(j)} = 0, \qquad l = 2, \cdots, n_j - 1,$

$(g_{n_j}^j)$ $\quad h_{n_j}^{(j)}x_{n_j}^{(j)} + b_j x_1^{(j+1)} = 0,$

with two exceptions:

(g_1^1) $\quad h_1^{(1)}x_1^{(1)} = 0,$

$(g_{n_k}^k)$ $\quad h_{n_k}^{(k)}x_{n_k}^{(k)} = 0.$

Suppose now (29), that some $\phi(A_m)$ satisfies (25) for interior entries. If $S_{m-1} = 0$, then $h_1^{(m)}$ is an interior entry, and as $x_{m-1} = 0$, (g_1^m) becomes $h_1^{(m)}x_1^{(m)} = 0$; if $S_{m-1} \neq 0$, then $h_1^{(m)}$ is not an interior entry (and similarly for $h_{n_m}^{(m)}$ and $(g_{n_m}^m)$). If $x_m \neq 0$, then $\omega(x_n)$ satisfies (24), and as in the proof of Lemma 6, $h_l^{(m)}x_l^{(m)} \neq 0$ for some l and $Hx \neq 0$, a contradiction; we must have $x_m = 0$. But this implies, if $m > 1$, by (g_1^m), that $x_{m-1}^{(m-1)} = 0$, and $x_{m-1} = 0$, and $x = 0$; if $m = 1$, by $(g_{n_1}^1)$, $x_1^{(2)} = 0$, and $x_2 = 0$, and $x = 0$.

Suppose instead (30), and that the transition minor $h_{n_m}^{(m)}h_1^{(m+1)} - b_m c_m$ is nonzero. As equations $(g_{n_m}^m)$ and (g_1^{m+1}) hold,

$$h_{n_m}^{(m)}x_{n_m}^{(m)} + b_m x_1^{(m+1)} = 0, \qquad c_m x_{n_m} + h_1^{(m+1)}x_1^{(m+1)} = 0,$$

we must have $x_{n_m}^{(m)} = x_1^{(m+1)} = 0$, which implies that $x_m = x_{m+1} = 0$, which implies that $x = 0$. We have shown that if any one of (29)–(31) holds for A, then for each positive diagonal E, the corresponding $FEAF^{-1} = H + S$ satisfies (18′) and is stable; A is D-stable.

Suppose instead that none of (29)–(31) holds. We must have at least one $S_j \neq 0$. Note that, as $A \in P_0^+$, transition entries, if any, are nonzero.

For each $S_j \neq 0$, let $\tilde{H}_j$ be defined as H_j, except that any transition entries are replaced by zero. Now $\phi(\tilde{H}_j) = (i_1, \cdots, i_p)$ satisfies (24). By Lemmas 3 and 4, there exist positive diagonal matrix E_j, nonsingular diagonal F_j, and vector $x_j \neq 0$, $\omega(x_j) = \phi(\tilde{H}_j)$, so that $B_j = F_j E_j S_j F_j^{-1}$ is skew-symmetric, with eigenvalue $\lambda = i$ and associated eigenvector x_j.

We define E to be the direct sum of the E_j, and define F to be a direct sum of appropriate nonzero scalar multiples of the F_j so that if $B = FESF^{-1}$, $|b_{i+1,i}| = |b_{i,i+1}|$ for all $i = 1, \cdots, n-1$. (This is true for $b_{i+1,i}b_{i,i+1} < 0$ as all $B_j \neq 0$ are skew-symmetric.) For simplicity of notation, we assume $B = H + S$, as in (26)–(28), now with each S_j either 1×1 and zero, or skew-symmetric, and each $b_j = c_j \neq 0$. We know $S_1 \neq 0$, and thus $x_1 \neq 0$.

If $S_2 = 0$, then $x_2 = 0$, $S_3 \neq 0$, and $x_3 \neq 0$. Now as $\omega(x_1) = \phi(H_1)$, $H_{11}x_1 = 0$, and equation (f_1) holds. Equation (f_2) becomes (in the notation of equations (g_l^j))

$$(g_1^2) \qquad c_1 x_{n_1}^{(1)} + b_2 x_1^{(3)} = 0,$$

and clearly we can replace the vector x_3 by a nonzero multiple of x_3 so that (g_1^2) is satisfied.

On the other hand, if $S_2 \neq 0$, then $\tilde{H}_1 \neq H_1$, and $h_{n_1}^{(1)}$ and $h_1^{(2)}$ are transition entries. As $\omega(x_1) = \phi(\tilde{H}_1)$, $H_{11}x_1 = 0$ except in the last position, i.e., (g_l^1) holds, $l = 1, \cdots, n_1 - 1$, and as the transition minor $h_{n_1}^{(1)}h_1^{(2)} - b_1c_2 = 0$, we may replace the vector x_2 by a nonzero multiple of x_2 so that

$$h_{n_1}^{(1)}x_{n_1}^{(1)} + b_1 x_1^{(2)} = 0, \qquad c_2 x_{n_1}^{(1)} + h_1^{(2)}x_1^{(2)} = 0.$$

Continuing this process (by induction) we obtain a vector $x \neq 0$ which is an eigenvector of S, and which is also in the null space of H. By Theorem 1, B is not stable (and consequently A is not D-stable). □

Remarks. We note that in Lemmas 4 and 6, $(i_1, \cdots, i_p) \neq \varnothing$ cannot satisfy (24) for $n < 5$. Thus, for $n \times n$ tridiagonal matrices satisfying (23), with $n < 5$, D-stability is equivalent to P_0^+. A "first" example, with $n = 5$, $\phi(A) = (3)$, follows. Let

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

$A \in P_0^+$. In the context of Theorem 1, let $G = I$, so that

$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad S = G^{-1}S = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 \end{bmatrix}.$$

Now $Sx = ix$ for $x^T = (1, i, 0, i, -1)$ while $Hx = 0$, so that x is an eigenvector of $G^{-1}S$ in the null space of H. This means according to Theorem 1 that A is not stable and thus not D-stable. The magnitudes—but not the signs—of the nonzero entries of A are inconsequential in this example.

Similarly, in Theorem 3, the conditions (29), (30), (31) can simultaneously fail only for $n \geqq 4$. For $n \times n$ irreducible tridiagonal matrices with $n < 4$, D-stability is equivalent to P_0^+. An example in which (29), (30), (31) all fail, with $n = 4$, follows.

Let

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix};$$

$A \in P_0^+$. Let $G = I$, so that

$$H = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad S = G^{-1}S = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}.$$

Now $Sy = iy$ for $y = (1, i, -i, 1)^T$ while $Hy = 0$, so that A is not stable, and thus not D-stable.

REFERENCES

[1] D. CARLSON AND B. N. DATTA, *On the effective computation of the inertia of a nonhermitian matrix*, Numer. Math., 33 (1979), pp. 315–322.

[2] D. CARLSON AND R. LOEWY, *On ranges of Lyapunov transformations*, Linear Algebra Appl., 8 (1974), pp. 237–248.

[3] D. CARLSON AND H. SCHNEIDER, *Inertia theorems: The semidefinite case*, J. Math. Anal. Appl., 6 (1963), pp. 430–446.

[4] C. T. CHEN, *A generalization of the inertia theorem*, SIAM J. Appl. Math., 25 (1973), pp. 158–161.

[5] B. N. DATTA, *Applications of Hankel matrices of Markov parameters to the solutions of the Routh–Hurwitz and Schur–Cohn problems*, J. Math. Anal. Appl., 68 (1979), pp. 276–290.

[6] ———, *On the Routh–Hurwitz–Fujiwara and the Schur–Cohn–Fujiwara theorems*, Linear Algebra Appl., 22 (1979), pp. 235–246.

[7] ———, *Stability and D-stability*, Linear Algebra Appl. 21 (1978), pp. 135–141.

[8] F. R. GANTMACHER, *The Theory of Matrices*, vol. I, Chelsea, New York, 1959.

[9] M. L. J. HAUTUS, *Controllability and observability conditions for linear autonomous systems*, Nederl. Akad. Wetensch. Proc. Ser. A, 72 (1979), pp. 443–448.

[10] C. R. JOHNSON, *Sufficient conditions for D-stability*, J. Econom. Theory, 9 (1974), pp. 53–62.

[11] ———, *Price stability in unions of markets*, Proc. NSF–CBMS Regional Conference on the Stability of Dynamical Systems, Marcel Dekker, New York, 1977, Chapter 10.

[12] A. LYAPUNOV, *Problème général de la stabilité du mouvement*, Comm. Soc. Math. Kharkov (1892, 93); Annals of Mathematical Studies, 17, Princeton Univ. Pr., Princeton, NJ, 1947.

[13] M. MARCUS AND H. MINC, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, 1964.

[14] A. OSTROWSKI AND H. SCHNEIDER, *Some theorems on the inertia of general matrices*, J. Math. Anal. Appl., 4 (1962), pp. 72–84.

[15] O. TAUSSKY, *A generalization of a theorem of Lyapunov*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 640–643.

[16] H. K. WIMMER, *Inertia theorems for matrices, controllability and linear vibrations*, Linear Algebra Appl., 8 (1974), pp. 337–343.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 3, September 1982

0196-5212/82/0303-0005 $01.00/0

COLORING BLOCK DESIGNS IS NP-COMPLETE*

CHARLES J. COLBOURN†‡, MARLENE J. COLBOURN,†
KEVIN T. PHELPS§ AND VOJTECH RÖDL¶

Abstract. Coloring partial Steiner triple systems is shown to be NP-complete. Together with an embedding technique of Lindner, this provides a short proof of the NP-completeness of coloring block designs.

1. Preliminaries. A *balanced incomplete block design* $B[k, \lambda; v]$ is a v-set V together with a collection B of k-subsets of V called *blocks*; each 2-subset of V occurs in exactly λ blocks of B. A *t-coloring* of a block design (V, B) is a mapping $k: v \to \{1, 2, \cdots, t\}$, so that there is no block $\{v_1, \cdots, v_k\}$ having $k(v_1) = k(v_2) = \cdots = k(v_k)$. A block design is *$t$-chromatic*, or has *chromatic number t* if it is t-colorable but not $(t-1)$-colorable. Previous research has studied designs with given chromatic number [1], [3], [8] in the general context of coloring hypergraphs [4], [5].

The purpose of this note is to present a short proof that deciding whether a block design is t-colorable is NP-complete. Thus, a characterization of designs with given chromatic number will likely not be "good" in the accepted sense [6].

2. Coloring partial STS. A *Steiner triple system* (STS) is a $B[3, 1; v]$ design; a partial STS is obtained by relaxing the constraints, so that every 2-subset of V appears in at most one block. In this section, we establish the preliminary result that

THEOREM 2.1. *Deciding whether a partial* STS *is t-colorable is* NP*-complete for any fixed $t \geqq 3$.*

In order to prove this theorem, we construct t-chromatic partial STS in which any t-coloring assigns a fixed pair of elements different colors.

LEMMA 2.2. *For each $t \geqq 2$, there is a t-chromatic partial* STS *for which any t-coloring assigns the same color to two fixed elements.*

Proof. There are $(t+1)$-chromatic STS for all $t \geqq 2$ [3]. Suppose P is a $(t+1)$-chromatic STS. A triple is said to be *critical* if its deletion lessens the chromatic number of the partial STS. Starting with any $(t+1)$-chromatic system, we delete blocks until one becomes critical. Call this partial STS P. Deleting a critical block from P produces a t-colorable partial STS P'. Any t-coloring of P' assigns the same color to the three elements forming the critical block of P, since otherwise the t-coloring of P' would also t-color P, which is in contradiction to our assumptions. □

LEMMA 2.3. *For each $t \geqq 2$, there exists a t-chromatic partial* STS *P and a fixed pair of elements $\{x, x'\}$ of P, such that any t-coloring of P assigns a different color to x and x'.*

Proof. Let P be a partial STS with chromatic number t, having the property that any t-coloring of P assigns the same color to two given elements x and y. Denote the element set of P by $Q \cup \{x\}$. Take two copies of P, one on $Q_1 \cup \{x\}$ and one on $Q_2 \cup \{x\}$—i.e., two copies intersecting only at x. Add a new element x' and include

* Received by the editors September 21, 1981.

† Department of Computational Science, University of Saskatchewan, Saskatoon, Saskatchewan, S7N 0W0, Canada.

‡ The work of this author was supported in part by Natural Science and Engineering Research Council of Canada under grant A5047.

§ School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

¶ FJFI, CVUT, Husova 5, Praha 1, Czechoslovakia.

the block $\{y_1, y_2, x'\}$. This partial STS is t-chromatic, and any t-coloring must assign the same color to x, y_1, and y_2. Then x' must be colored differently from x.

Proof of Theorem 2.1. Suppose we are to decide whether an arbitrary graph G is t-colorable; we know that this problem is NP-complete for any fixed $t \geqq 3$ [6]. First, let P be a partial STS with chromatic number t having fixed elements x, x' which every t-coloring of P assigns two different colors. We construct a partial STS with a copy of P for every edge of the graph G; for an edge $\{y, z\}$ of G, we take a copy of P disjoint from the other copies, and identify x and x' with y and z. The theorem follows directly. □

3. Coloring block designs. We prove in this section that

THEOREM 3.1. *Deciding whether a block design is t-colorable is* NP*-complete for all* $t >= 9$.

Proof. Membership of this coloring problem in NP is immediate. Thus we need only provide a polynomial time reduction of a known NP-complete problem to our coloring problem. In light of Theorem 2.1, it suffices to show that, given a partial STS on v elements, we can produce in polynomial time a $B[3, 12tv+3; 18tv+3]$ which is $3t$-colorable if and only if the partial STS is t-colorable.

Commencing with a partial STS P on v elements, we first take $3t$ disjoint copies of P to form P'. Using Cruse's method [2], we produce in polynomial time a commutative idempotent quasigroup of order $6tv+1$ which contains the partial Steiner quasigroup corresponding to P'. Using a construction of Lindner [7], we next produce an STS of order $18tv+3$ as follows. Let $s = 6tv+1$. Our STS has element set $\{x_1, \cdots, x_s, y_1, \cdots, y_s, z_1, \cdots, z_s\}$.

We process each pair $\{a, b\}$ for $1 \leqq a < b \leqq s$ in turn. If $\{a, b\}$ belongs to a triple of P', say the triple $\{a, b, c\}$, our STS contains the nine blocks $\{x_a, y_b, z_c\}$, $\{x_a, y_c, z_b\}$, $\{x_b, y_a, z_c\}$, $\{x_b, y_c, z_a\}$, $\{x_c, y_a, z_b\}$, $\{x_c, y_b, z_a\}$, $\{x_a, x_b, x_c\}$, $\{y_a, y_b, y_c\}$, and $\{z_a, z_b, z_c\}$. On the other hand, if $\{a, b\}$ does not appear in a block of P', we look up $ab = c$ in the commutative quasigroup and add the three blocks $\{x_a, x_b, y_c\}$, $\{y_a, y_b, z_c\}$, and $\{z_a, z_b, x_c\}$. Finally, we add the block $\{x_i, y_i, z_i\}$ for each $1 \leqq i \leqq s$. This is an STS of order $18tv+3$.

We transform this into a $B[3, 12tv+3; 18tv+3]$ B by adding the following blocks:

(1) $\{x_i, x_j, y_k\}$, $\{x_i, x_j, z_k\}$, $1 \leqq i < j \leqq s$, $1 \leqq k \leqq s$,

(2) $\{y_i, y_j, x_k\}$, $\{y_i, y_j, z_k\}$, $1 \leqq i < j \leqq s$, $1 \leqq k \leqq s$,

(3) $\{z_i, z_j, x_k\}$, $\{z_i, z_j, y_k\}$, $1 \leqq i < j \leqq s$, $1 \leqq k \leqq s$,

and

(4) $\{x_i, y_{i+m}, z_{i+2m}\}$, $1 \leqq i \leqq s$, $1 \leqq m \leqq s$, each included twice.

In (4), subscripts are reduced into range as required. This collection of blocks is a block design. Moreover, if P' is t-colorable, B is $3t$-colorable; a $3t$-coloring uses t colors for each "level", i.e., for each of the $\{x_i\}$, the $\{y_i\}$ and the $\{z_i\}$. Within a level, the t colors form a t-coloring of P', and any of the t colors can be used for the elements of the level not in the copy of P'.

We claim also that if B is $3t$-colorable, P' is t-colorable. If the colors assigned to each level are disjoint, the t colors on a level induce a t-coloring of P'. Otherwise, remark that a color appearing on more than one level appears exactly once on each level as a consequence of the blocks (1)–(3). But then one of the $3t$ copies of P must fail to contain an element with such a color, and hence must have colors used only on that level; hence, P is t-colored.

This completes the polynomial time reduction as required. □

The result of Theorem 3.1 is not unexpected. Its importance derives in part from the relative simplicity of the proof; more important, however, is the fact that few algorithmic problems in design theory are known to be NP-complete, despite expectations that many are. Thus Theorem 3.1 provides a first example of such a problem, and supplies a building block for proving further NP-completeness results in computational design theory.

REFERENCES

[1] C. J. COLBOURN, M. J. COLBOURN, K. T. PHELPS AND V. RÖDL, *Colouring Steiner quadruple systems*, Discrete Appl. Math., to appear.

[2] A. B. CRUSE, *On embedding incomplete symmetric Latin squares*, J. Combin. Theory Ser. A., 16 (1974), pp. 18–22.

[3] M. DE BRANDES, K. T. PHELPS AND V. RÖDL, *Coloring Steiner triple systems*, this Journal, 3 (1982), pp. 241–249.

[4] P. ERDÖS AND A. HAJNAL, *On the chromatic number of graphs and set systems*, Acta Math. Acad. Sci. Hungar., 17 (1966), pp. 61–99.

[5] P. ERDÖS AND L. LOVÁSZ, *Problems and results on 3-chromatic hypergraphs and related questions*, in Infinite and Finite Sets, North-Holland, Amsterdam, 1975, pp. 609–627.

[6] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability*, W. H. Freeman, San Francisco, 1979.

[7] C. C. LINDNER, *A partial Steiner triple system can be embedded in a Steiner triple system of order* $6n+3$, J. Combin. Theory Ser. A, 18 (1975), pp. 349–351.

[8] A. ROSA, *Steiner triple systems and their chromatic number*, Acta Fac. Rerum Natur. Univ. Comenian. Math., 24 (1970), pp. 159–174.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 3, September 1982

0196-5212/82/0303-0006 $01.00/0

SPREADS, TRANSLATION PLANES AND KERDOCK SETS. II*

W. M. KANTOR†

Abstract. New Kerdock sets of q^{2n-1} skew-symmetric $2n \times 2n$ matrices over $GF(q)$ are constructed for even q whenever $2n-1$ is composite. Related affine translation planes are studied in detail. In both cases, explicit coordinate descriptions are given.

1. Introduction. This paper is a continuation of [6], hereafter called [STK]. In that paper, the relationship between spreads, translation planes and Kerdock sets was described. Nondesarguesian examples were given, arising either from slices of desarguesian spreads or from certain spreads in $\Omega^+(8, q)$ spaces. In this paper we will study these slices more closely, and construct new Kerdock sets in higher dimensional spaces.

When an $\Omega^+(2m, q^e)$ space is turned into an $\Omega^+(2em, q)$ space by following the quadratic form with the trace map, new singular vectors are introduced. Thus, it is not possible to directly change the dimension of the space in which an orthogonal spread lies in order to obtain a new spread. However, when an $Sp(2m, q^e)$ space is turned into an $Sp(2em, q)$ space, this difficulty does not arise. This produces the following construction (§ 2): take an orthogonal spread, slice in order to obtain a symplectic spread, change fields, and then embed the resulting symplectic spread as a slice of a new orthogonal spread.

This procedure provides us with a machine for grinding out large numbers of new spreads and new translation planes. These do not seem to have new properties: from the point of view of their groups, they have fewer properties than the original spreads and planes. On the other hand, the procedure requires a great deal of interesting interplay between orthogonal spreads and translation planes. It is not at all clear how one can directly pass from an orthogonal spread to one of the many new ones it spawns; it seems even less likely that one could directly pass from a Kerdock set over $GF(q^e)$ to one of the many new ones over $GF(q)$.

The spreads and Kerdock sets obtained in this manner from the unitary spreads of [STK, § 6] are new for trivial reasons (§ 3), but are difficult to compute with.

Most of the paper is devoted to spreads obtained by starting with the desarguesian plane $AG(2, (q^e)^m)$, passing to one of its "cousins", and then changing fields. The resulting orthogonal spreads and Kerdock sets are studied in §§ 8 and 9. In order to show that these are new, we must study the aforementioned cousins rather carefully. This is done by using their coordinatizing quasifields in §§ 5 and 6.

The spreads of $\Omega^+(2m, 2)$ spaces obtained here produce partial geometries as in DeClerck, Dye and Thas [2], having the same parameters as their partial geometries but not isomorphic to their "desarguesian" ones.

2. Expanding spreads: definition. Let q be a power of 2, let m be odd with $m > 1$, and let Σ be a spread of an $\Omega^+(2m+2, q^e)$ space V. If e is odd, then many spreads can be constructed in $\Omega^+(2em+2, q)$ spaces, as follows.

Let y be any nonsingular point of V, and form the spread

$$\Sigma(y) = (y^\perp \cap \Sigma)/y = \{\langle y, y^\perp \cap F\rangle/y \mid F \in \Sigma\}$$

* Received by the editors July 14, 1981.

† Bell Laboratories, Murray Hill, New Jersey 07974. Permanent address: Mathematics Department, University of Oregon, Eugene, Oregon 97403.

in the $Sp(2m, q^e)$ space $y^\perp/y$ [STK, (3.1)]. If $(\cdot,\cdot)$ is the symplectic form on $y^\perp/y$, and $T: GF(q^e) \to GF(q)$ is the trace map, then $T(\cdot,\cdot)$ turns $y^\perp/y$ into an $Sp(2em, q)$ space. Totally isotropic spaces remain totally isotropic. Thus, $\Sigma(y)$ produces a spread $\Sigma(y)^e$ of totally isotropic em-spaces of $y^\perp/y$. (Of course, $\Sigma(y)$ and $\Sigma(y)^e$ determine the same translation plane $\mathbf{A}(\Sigma(y))$.) Finally, form the spread $\mathbf{S}(\Sigma(y)^e)$ of an $\Omega^+(2em+2, q)$ space, as in [STK, (3.2)]. Note that e must be odd here, in order to have $2em+2 \equiv 0 \pmod 4$.

The procedure described above will be called *expanding* the spread $\Sigma(y)$ into $2em+2$ dimensions.

In view of [STK, § 3], $\mathbf{S}(\Sigma(y)^e)$ determines many translation planes defined by symplectic spreads $(z^\perp \cap \mathbf{S}(\Sigma(y)^e))/z$, where z is any nonsingular point of the various $\Omega^+(2em+2, q)$ spaces. $\mathbf{S}(\Sigma(y)^e)$ also determines many Kerdock sets of $(em+1)\times(em+1)$ skew-symmetric matrices over $GF(q)$, as in [STK, § 5].

The remainder of this paper will be concerned with examples of such expanded spreads, planes and Kerdock sets are new.

Of crucial importance are the following trivial observations.

LEMMA 2.1. *Let Σ, y and $\Sigma^* = \mathbf{S}(\Sigma(y)^e)$ be as above.*

(i) *There is a nonsingular point y^* of the underlying $\Omega^+(2em+2, q)$ space such that $\Sigma(y)^e = \Sigma^*(y^*)$.*

(ii) *$\Gamma O^+(2m+2, q^e)_{\Sigma, y}$ induces a subgroup of $\Gamma O^+(2em+2, q)_{\Sigma^*, y^*}$ in such a way that the permutation representations on Σ and Σ^* are equivalent.*

3. Unitary spreads. The expanded examples which are most easily shown to be new arise from the unitary spreads constructed in [STK, § 6]. Such a spread Σ arises in an $\Omega^+(8, q^e)$ space, where $\log_2 q^e$ is odd and $q^e > 2$. Define N and M as in [STK, Thm. 7.1, Example 7.5]. Assume that $e > 1$.

THEOREM 3.1. *The expanded spreads $\mathbf{S}(\Sigma(\langle N\rangle)^e)$ and $\mathbf{S}(\Sigma(\langle M\rangle)^e)$ are nondesarguesian spreads in $\Omega^+(6e+2, q)$ space.*

Proof. There is a subgroup $G = PGU(3, q^3)$ of $\Gamma O^+(8, q^e)_\Sigma$. By [STK, § 7], $G_N = GU(2, q^e)$ has a cyclic normal subgroup of order q^e+1 fixing q^e+1 members of Σ. That cyclic group acts on $\mathbf{S}(\Sigma(\langle M\rangle)^e)$ by Lemma 2.1. However, the subgroup of $\Gamma O^+(6e+2, q)$ preserving a desarguesian spread induces $P\Gamma L(2, q^{3e})$ on that spread [STK, (4.1)], and hence cannot have a cyclic subgroup acting as above. Consequently, $\mathbf{S}(\Sigma(\langle N\rangle)^e)$ is nondesarguesian, and the same argument shows that $\mathbf{S}(\Sigma(\langle M\rangle)^e)$ also is.

THEOREM 3.2. *$\mathbf{S}(\Sigma(\langle N\rangle)^e)$ and $\mathbf{S}(\Sigma(\langle M\rangle)^e)$ are not equivalent under the action of $\Gamma O^+(6e+2, q)$.*

Proof. This requires some group theory, and will only be briefly sketched. Assume that these expanded spreads are equivalent. Call either of them Σ^*. Then $H = \Gamma O^+(6e+2, q)_{\Sigma^*}$ contains subgroups acting on Σ^* as G_N and G_M do on their respective symplectic spreads. It follows that H acts transitively on Σ^*. A detailed analysis yields that H acts 2-transitively on Σ^*. However, this is impossible in view of the following lemma. □

LEMMA 3.3. *Let Σ be a spread in an $\Omega^+(2n+2, q)$ space V, where q is even, n is odd and $n > 3$. Assume that $\Gamma O^+(2n+2, q)_\Sigma$ is 2-transitive on Σ. Then Σ is desarguesian.*

Proof. First note that q^n+1 is not a prime power. Then $\Gamma O^+(2n+2, q)_\Sigma$ has a subgroup G inducing $PSL(2, q^n)$ or $PSU(3, q^{n/3})$ on Σ (Holt [5, Thm. 2]). Here, G has a $GF(2)$-representation on our space of size q^{2n+2}. It follows that G cannot act irreducibly on V, and fixes some 1-space z (Fong and Seitz [4, (4A), (4B), (4D)]). Clearly, z cannot be singular. Thus, G acts on the translation plane $\mathbf{A}(\Sigma(z))$, inducing

$PSL(2, q^n)$ or $PSU(3, q^{n/3})$ on the line at infinity. Consequently, $\Sigma(z)$ is desarguesian (Lüneburg [7, pp. 178–179]), and hence so is Σ (by definition [STK, § 4]). □

There are many translation planes arising from the spreads in Theorem 3.1 [STK, § 3], but none seems manageable or interesting.

4. Desarguesian spreads. In order to deal with the expanded cousins of desarguesian spreads, we will need to study these cousins using their coordinatizing quasifields. This in turn requires a description of the corresponding spreads, and hence of desarguesian spreads of $\Omega^+(2m+2, q)$ spaces.

Let q be even and m be odd. Set $F = GF(q^m)$ and $K = GF(q)$ throughout §§ 4–7. Let $T: F \to K$ be the trace map. Then T is K-linear, and satisfies

$$T(\alpha)^2 = T(\alpha^2) \quad \text{for all } \alpha \in F, \qquad T(a) = a \quad \text{if } a \in K. \tag{4.1}$$

Let V_0 be the F-space with basis e, f, view V_0 as a $2m$-dimensional K-space, and form the $(2m+2)$-dimensional space $V = V_0 \oplus \langle u, w \rangle$. Define a quadratic form Q on V by

$$Q(\alpha e + \beta f + cu + dw) = T(\alpha\beta) + c^2 + cd.$$

This turns V into an $\Omega^+(2m+2, q)$ space.

Set

$$\begin{aligned} \Sigma[\infty] &= Ff + K(u+w), \\ \Sigma[s] &= \{\alpha e + (s^2\alpha + sa)f + T(s\alpha)u + aw \mid \alpha \in F, a \in K\} \end{aligned} \tag{4.2}$$

for $s \in F$. Then

$$\Sigma = \{\Sigma[s] \mid s \in F \cup \{\infty\}\}$$

is a desarguesian spread in V.

Define linear transformations j and $[t]$ as follows (where $t \in F$).

$$j: \begin{cases} \alpha e \leftrightarrow \alpha f \\ u \to u \\ w \to w + u \end{cases} \qquad [t]: \begin{cases} \alpha e \to \alpha e + \alpha t^2 f + T(\alpha t)u \\ f \to f \\ u \to u \\ w \to w + tf \end{cases} \tag{4.3}$$

Then j and $[t]$ preserve Q, and act on Σ as follows: $\Sigma[s]^j = \Sigma[s^{-1}]$ and $\Sigma[s]^{[t]} = \Sigma[s+t]$. Thus, $G = \langle j, [t] \mid t \in F \rangle$ induces $SL(2, q^m)$ on Σ, and is, in fact, isomorphic to $SL(2, q^m)$. The action of G (and even of $P\Gamma L(2, q^m)$) on Σ is used in [STK, § 4] in order to distinguish between the various cousins of $AG(2, q^m)$.

Every cousin has the form $\Sigma(y)$, with $y = \langle u \rangle, \langle f+u \rangle, \langle u + kw \rangle$ with $k \in K - GF(2)$, or $\langle ku + w + r(e+f) \rangle$ with $k \in K$, $r \in K$ and $x^2 + x + r$ irreducible. These cousins are, respectively, the first, second, third and fourth cousins of $AG(2, q^m)$ [STK, Thm. 4.2].

$\Sigma(\langle u \rangle)$ produces $AG(2, q^m)$.

Second cousin.

$$\begin{aligned} &\Sigma(\langle f+u \rangle)[\infty] = Ff, \\ &\Sigma(\langle f+u \rangle)[s] = \{\alpha e + (s^2\alpha + sT(\alpha) + T(s\alpha))f \mid \alpha \in F\}, \end{aligned} \tag{4.4}$$

Third cousins.

$$\begin{aligned} &\Sigma(\langle u + kw \rangle)[\infty] = Ff, \\ &\Sigma(\langle u + kw \rangle)[s] = \{\alpha e + (s^2\alpha + ksT(s\alpha))f \mid \alpha \in F\}. \end{aligned} \tag{4.5}$$

In the above spreads, we have projected onto V_0 in order to obtain these relatively simple descriptions. Note that (4.4) and (4.5) are both symplectic relative to the natural symplectic form $(\alpha e+\beta f, \alpha' e+\beta' f)=T(\alpha\beta'+\alpha'\beta)$ on V_0.

Fourth cousins are also easily computed as $\Sigma(\langle ku+r(e+f)\rangle)$. However, the resulting spreads and quasifields seem difficult to compute with. An alternative description of fourth cousins will be used in § 7.

5. Second cousins. The next two sections consist of coordinate calculations with second and third cousins of desarguesian planes. These calculations are needed for Lemma 7.1, which is a crucial step in our proof (in Theorem 8.1) that expansions of these cousins are new. For completeness, we will provide an alternative verification of the fact that these cousins are nondesarguesian [STK, Thm. 4.2].

Let F, K and T be as in § 4. The spread (4.4) yields a semifield, which we now proceed to describe.

For $x, y \in F$, write

$$x * y = x^2 y + xT(y) + T(xy).$$

LEMMA 5.1. *If $x * y = 0$ then $x = 0$ or $y = 0$.*

Proof. Write $z = xy$. Then $z^2 + zT(y) + yT(z) = 0$. Apply T and obtain $T(z)^2 + T(z)T(y) + T(y)T(z) = 0$. Then $T(z) = 0$ and $z^2 = zT(y)$. If $z \neq 0$ then $z = T(y)$, so that $0 = T(z) = T(y)$ (by (4.1)). Thus, $z = 0$. □

DEFINITION 1. $\bar{x}$ is the unique solution to

$$\bar{x}^2 + \bar{x} + T(\bar{x}) = x. \tag{5.2}$$

Thus, $x \to \bar{x}$ is the inverse of the map $x \to x * 1$. Note that $\bar{a} = a^{1/2}$ if $a \in K$, while $T(\bar{x})^2 = T(x)$ by (4.1).

DEFINITION 2. $x \circ y = \bar{x} * y = \bar{x}^2 y + \bar{x} T(y) + T(\bar{x}y)$.

THEOREM 5.3. *$(F, \circ)$ is a semifield. It is not a field if $q^m > 8$.*

Proof. If $a \in K$ then $a \circ x = ax = x \circ a$. Also, $x \to \bar{x}$ is additive. By Lemma 5.1, $(F, \circ)$ is a semifield. The second part of Theorem 5.3 follows from the next two lemmas.

LEMMA 5.4. *$GF(2) = \{a \in K \mid (a \circ u) \circ v = a \circ (u \circ v)$ for all $u, v \in F\}$.*

Proof. Assume that $(a \circ u) \circ v = a \circ (u \circ v)$ for all $u, v \in F$, where $a \in K - GF(2)$. Since $(a \circ u) \circ v = (au) \circ v$ and $a \circ (u \circ v) = a(u \circ v)$, we have

$$\overline{au}^2 v + \overline{au} T(v) + T(\overline{au}v) = a(\bar{u}^2 v + \bar{u} T(v) + T(\bar{u}v)),$$

$$(\overline{au}^2 + a\bar{u}^2)v = (\overline{au} + a\bar{u})T(v) + (T(\overline{au}v) + aT(\bar{u}v)).$$

If $\overline{au}^2 + a\bar{u}^2 \neq 0$ for some u, we can divide in order to obtain $\dim_K F \leqq 2$. Thus, $\overline{au}^2 + a\bar{u}^2 = 0$ for all u.

From (5.2) it now follows that

$$\overline{au} + au + T(\overline{au}) = a(\bar{u} + u + T(\bar{u}))$$

or

$$\overline{au} + a\bar{u} = T(\overline{au} + a\bar{u})$$

for all $u \in F$. Since $\overline{au}^2 = a\bar{u}^2$, $a^{1/2}\bar{u} + a\bar{u} \in K$ for all $u \in F$. However, $a^{1/2} + a \neq 0$, so this is impossible. □

LEMMA 5.5. *If $q^m > 8$ then $K = \{z \in F \mid (u \circ v) \circ z = u \circ (v \circ z)$ for all $u, v \in F\}$.*

Proof. Call the indicated set L. If $a \in K$ then, by definition, $(u \circ v) \circ a = (u \circ v)a$ and $u \circ (v \circ a) = \bar{u}^2(va) + \bar{u} T(va) + T(\bar{u}(va))$. Thus, $L \supseteq K$.

Note that $(L, \circ)$ is a field. The maps $u \to u \circ z$ for $z \in L^*$ form a group acting semiregularly on F^*. If L is $GF(q^l)$, then $q^l - 1 | q^m - 1$, so $l | m$. Assume that $l > 1$. Since m is odd, $l \geqq 3$. Let $a \in K - GF(2)$. Then $(a \circ u) \circ v = a \circ (u \circ v)$ for all u, v in the field L. The argument in Lemma 5.4 can now be repeated (with u and v always in L) in order to obtain a contradiction. This completes the proof of both Lemma 5.5 and Theorem 5.3. □

6. Third cousins. Let F, K and T be as in § 4, with $q > 2$. Fix $k \in K - GF(2)$. Using (4.5), we will again define $x * y$, $\bar{x}$ and $x \circ y$; however, these expressions will have nothing to do with those of the preceding section (except, of course, for the fact that the corresponding planes are cousins).

For $x, y \in F$ write

$$x * y = x^2 y + kxT(xy).$$

LEMMA 6.1. *If* $u * y - v * y = 0$ *then* $u = v$ *or* $y = 0$.

Proof. If $(u * y - v * y)y = 0$ then

$$u^2 y^2 + kuyT(uy) = v^2 y^2 + kvyT(vy).$$

Set $\alpha = uy$ and $\beta = vy$. Then

$$\alpha^2 + k\alpha T(\alpha) = \beta^2 + k\beta T(\beta).$$

Apply T, and use (4.1):

$$T(\alpha)^2 + kT(\alpha)T(\alpha) = T(\beta)^2 + kT(\beta)T(\beta).$$

Thus, $T(\alpha) = T(\beta)$, so $\alpha^2 + \beta^2 = k(\alpha + \beta)T(\alpha)$. If $\alpha = \beta$, the lemma is clear. If $\alpha \neq \beta$ then $\alpha + \beta = kT(\alpha)$, so $0 = T(\alpha) + T(\beta) = kT(\alpha)$ and $\alpha + \beta = 0$. Thus, $\alpha = \beta$, as required. □

DEFINITION 3. Let $x \to \bar{x}$ be the inverse of the map $x \to (x * 1)/(k + 1)$. Thus,

$$x = \frac{\bar{x}^2 + k\bar{x}T(\bar{x})}{k+1}. \tag{6.2}$$

Apply T and obtain $T(x) = T(\bar{x})^2$. Also, $\bar{a} = a^{1/2}$ if $a \in K$.

DEFINITION 4. Let $y \to y'$ be the inverse of $y \to (1 * y)/(k + 1)$. Thus,

$$y = \frac{y' + kT(y')}{k+1}.$$

This time, $T(y) = T(y')$, and we can solve for y':

$$y' = (k+1)y + kT(y).$$

Then $a' = a$ if $a \in K$.

DEFINITION 5. $x \circ y = (\bar{x} * y')/(k + 1)$.

THEOREM 6.3. $(F, \circ)$ *is a quasifield. It is never a field.*

Proof. If $a \in K$ then $a \circ y = (\bar{a} * y')/(k+1) = (a^{1/2} * y')/(k+1) = (ay' + kaT(y'))/(k+1) = y$ and $x \circ a = (\bar{x} * a')/(k+1) = (\bar{x} * a)/(k+1) = a(\bar{x} * 1)/(k+1) = ax$. By Lemma 6.1, $(F, \circ)$ is thus a quasifield. The theorem is then a consequence of the next result.

LEMMA 6.4. $K = \{y \in F | (u + v) \circ y = u \circ y + v \circ y$ *for all* $u, v \in F\}$.

Proof. Let L denote the right-hand set. If $a \in K$ then $u \circ a = ua = a \circ u$, so that $(u + v) \circ a = u \circ a + v \circ a$ and $K \subseteq L$. Note that $y \in L$ if and only if $\overline{u+v} * y' = \bar{u} * y' + \bar{v} * y'$ for all $u, v \in F$. Set $L' = \{y' | y \in L\}$. Then L' consists of all $\zeta \in F$ such

that the following holds for all $u, v \in F$:

$$\overline{u+v}^2\zeta + k\overline{u+v}T(\overline{u+v}\zeta) = \bar{u}^2\zeta + k\bar{u}T(\bar{u}\zeta) + \bar{v}^2\zeta + k\bar{v}T(\bar{v}\zeta). \tag{6.5}$$

Thus, L' is a vector space over K, and $L' \supseteq K$. We must show that $L' = K$.

Assume that $\dim L' \geqq 2$. Define a nonsingular symmetric K-bilinear form on F by setting $(x, y) = T(xy)$. Then $1^\perp$ is the space of trace 0 elements F, and $1^\perp \supset L'^\perp$. From now on, $\bar{u}$ and $\bar{v}$ will be chosen from $1^\perp$. Since $T(\alpha) = T(\bar{\alpha})^2$, this amounts to choosing $u, v \in 1^\perp$. Then $u+v$ and $\overline{u+v}$ also belong to $1^\perp$.

By (6.2), $\bar{u}^2 = (k+1)u$. Thus, (6.5) reduces to

$$(u+v)T((u+v)\zeta^2) = uT(u\zeta^2) + vT(v\zeta^2)$$

for all $u, v \in 1^\perp$ and $\zeta \in L'$. Then $uT(v\zeta^2) = vT(u\zeta^2)$. Since $\dim 1^\perp > \dim L'^\perp$ we can find $v \in 1^\perp$ and $\zeta \in L'$ such that $T(v\zeta^2) = (v, \zeta^2) \neq 0$. Then each $u \in 1^\perp$ lies in the 1-space Kv. Since $\dim 1^\perp = m-1 \geqq 2$, this is ridiculous. This completes the proof of both Lemma 6.4 and Theorem 6.3. □

By definition, the plane over $(F, \circ)$ has a very nice collineation of order $q^m - 1$ [STK, Thm. 4.2(iii)]. For completeness, we will exhibit this collineation.

PROPOSITION 6.6. *There is a collineation g of order $q^m - 1$ which fixes 0 and two points x_∞ and y_∞ at infinity, such that $\langle g \rangle$ has orbits of length $q^m - 1$ on the lines $0x_\infty$, $0y_\infty$ and $x_\infty y_\infty$.*

Proof. The lines through the origin of the plane over $(F, \circ)$ are $x = 0$ and $y = n \circ x$. Define g by

$$(x, y)^g = ([1^*(\zeta^{-1}x')]/(k+1), \zeta y),$$

where $\langle \zeta \rangle = GF(q^m)^*$. Clearly, $x = 0$ and $y = 0$ are fixed lines. We will show that g sends $y = n \circ x$ to $y = r \circ x$, where $\bar{n} = \bar{r}\zeta$.

By definition, $(x, n \circ x)^g = (u, \zeta(n \circ x))$, where $u = [1 * (\zeta^{-1}x')]/(k+1)$. The definition of u' shows that $u' = \zeta^{-1}x'$, so $x' = \zeta u'$. Now

$$\begin{aligned} \zeta(m \circ x) &= \zeta(\bar{m} * (\zeta u'))/(k+1) \\ &= \zeta[\bar{m}^2\zeta u' + k\bar{m}T(\bar{m}\zeta u')]/(k+1) \\ &= (\bar{r} * u')/(k+1) = r \circ u. \end{aligned}$$

Thus, g sends points of the form $(x, m \circ x)$ to points of the form $(u, r \circ u)$.

Since both $x \to (x * 1)/(k+1)$ and its inverse $x \to x'$ are additive (in fact, K-linear), g is a collineation. Moreover, the relations $\bar{r} = \bar{n}\zeta$ and $u' = x'\zeta^{-1}$ prove the desired transitivity on the line $y = 0$ and the line at infinity; on the line $x = 0$, this transitivity is obvious. This proves the result. □

Remarks. 1. Let $F(k)$ denote the quasifield in Theorem 6.3. Clearly, $\mathrm{Gal}(GF(q^m)/GF(q))$ lies in $\mathrm{Aut}\, F(k)$, while $\mathrm{Aut}\, GF(q^m)$ does not.

If $q = p^2$ is a square, and $k^p = k$, then the involutory field automorphism θ defined by $x^\theta = x^{p^m}$ is in $\mathrm{Aut}\, F(k)$. If $x^\theta = x$ then $T(x)$ is also obtained from the trace map $GF(p^m) \to GF(p)$, and we obtain a Baer subplane which can be coordinatized by a quasifield obtained in the same manner as $F(k)$ was. In particular, this subplane is nondesarguesian.

2. In the notation of (4.5), $F(k)$ arises from the nonsingular point $\langle u + kw \rangle$. By (4.3), $\langle u + kw \rangle^j = \langle u + k_1 w \rangle$ with $k_1 = k/(k+1)$, so the planes over $F(k)$ and $F(k_1)$ are

isomorphic. This accounts for the 2 in the denominator occurring in [STK, Thm. 4.2(iii)].

3. By [STK, Thm. 4.2(iii)], the plane over $F(k)$ is not a semifield plane. It seems to be difficult to prove this directly from the definition of $F(k)$.

7. Homologies. Let **A** be a second, third or fourth cousin of $AG(2, q^m)$. The group induced by Aut (**A**) on the line at infinity is described in [STK, Thm. 4.2]. In order to deal with expansions of these planes, we will need information concerning the groups of homologies with center 0. This amounts to an easy application of parts of the last two sections when **A** is a second or third cousin; however, a different approach is required in order to prove the corresponding result for fourth cousins.

LEMMA 7.1. *Let* **A** *be a second, third or fourth cousin of* $AG(2, q^m)$, *where* $q^m > 8$. *Let* H *denote the group of all homologies with center* 0. *Then* $H \cong GF(q)^*$.

Proof. If $(F, \circ)$ is one of the quasifields in §§ 5 or 6, then H is isomorphic to the group of all $x \in F^*$ such that $(u+v)\circ x = u\circ x + v\circ x$ and $(u\circ v)\circ x = u\circ(v\circ x)$ for all $u, v \in F$ (Dembowski [3, p. 132]). Now apply Lemmas 5.5 and 6.4.

The remainder of this section will be devoted to the case of fourth cousins. In order to prove Lemma 7.1 in this case, we will need a description of their spreads. This will be obtained from a description of desarguesian spreads different from that of § 4.

Let K, F and T be as usual. Let $F' = GF(q^{2m})$ and $K' = GF(q^2)$. We will depart from the notation in § 4 by writing $V = F' \oplus K'$ and

$$Q(\alpha, r) = T(\alpha\bar{\alpha}) + r\bar{r}$$

for $\alpha \in F'$ and $r \in K'$; here, $\bar{\alpha} = \alpha^{q^m}$. Note that $K' \not\subseteq F$.

Let W denote the kernel of T. Set

$$\Sigma[\theta] = \{(\theta w + \theta r, r) | w \in W, r \in K'\}$$

whenever $\theta\bar{\theta} = 1$, and

$$\Sigma = \{\Sigma[\theta] | \theta\bar{\theta} = 1\}.$$

Then $\Sigma[\theta]$ is a totally singular $m+1$-space (so that V is an $\Omega^+(2m+2, q)$ space), and Σ is a spread.

If $r \neq 0$ then $(0, r)$ is nonsingular. Set $\Sigma' = \Sigma(((0, r)))$. Then Σ' is a symplectic spread, which can be identified with the set of all K-subspaces

$$\Sigma'[\theta] = \theta W + \theta r K \tag{7.2}$$

of F', where $\theta\bar{\theta} = 1$.

If $r \in K$ then (7.2) states that $\Sigma'[\theta] = \theta F$. Thus, Σ' is desarguesian in this case, and hence so is Σ.

Every $r \in K' - K$ determines a fourth cousin of $AG(2, q^m)$. Clearly, r and ar determine the same cousin if $a \in K^*$. Note that r and $\bar{r}$ determine isomorphic cousins (compare [STK, Thm. 4.2(iv)]).

If $\phi\bar{\phi} = 1$, then $x \to \phi x$ sends $\Sigma'[\theta]$ to $\Sigma'[\phi\theta]$. This produces the cyclic collineation group appearing in [STK, Thm. 4.2(iv)].

We are now in a position to complete the proof of Lemma 7.1. Fix $r \in K' - K$. The group H of homologies of $\mathbf{A}(\Sigma')$ with center 0 consists of those invertible semilinear transformations of the K-space F' which induce the identity on Σ'.

Assume that $|H| > q-1$. Clearly, H is normalized by the above cyclic collineation group of order q^m+1. It follows that there is an irreducible collineation group $\langle g \rangle$

centralizing some $h \in H$ such that $|h| \nmid q-1$. By Schur's lemma, $C_{\Gamma L(F')}(g) \cong F'^*$. Consequently, h has the form $x \to lx$ for some $l \in F'-K$.

Now $l\Sigma'[1]=\Sigma'[1]$. Since $\dim W = m-1 \geqq 2$, $lW \cap W \neq 0$. Then $l \in F$, so that $lW \subseteq F \cap \Sigma'[1] = W$ (since $r \in K'-K$). Consequently, $|l|$ divides both $|F|-1$ and $|W|-1$. However, $|l| \nmid q-1$, so this is ridiculous.

This completes the proof of Lemma 7.1. □

Note that the above argument provides a direct verification of the fact that fourth cousins are nondesarguesian.

8. Expanded cousins of desarguesian spreads. We are finally ready to deal with the spreads $\mathbf{S}(\Sigma(y)^e)$ obtained from a desarguesian spread Σ of an $\Omega^+(2m+2, q^e)$ space, where e and m are odd.

THEOREM 8.1. *If $\Sigma(y)$ is a second, third or fourth cousin of the $AG(2,(q^e)^m)$ spread, where $e>1$, $m>1$ and em is odd, then $\mathbf{S}(\Sigma(y)^e)$ is a nondesarguesian spread in an $\Omega^+(2em+2, q)$ space.*

Proof. Assume that $\Sigma^* = \mathbf{S}(\Sigma(y)^e)$ is desarguesian. Let y^* be as in Lemma 2.1. Then $\Sigma^*(y^*)$ must be a cousin of $AG(2, q^{em})$, while $\Sigma^*(y^*) = \Sigma(y)^e$. Thus, $\Sigma^*(y^*)$ is nondesarguesian. Now two applications of Lemma 7.1 produce a contradiction. □

THEOREM 8.2. (i) *The nondesarguesian spreads in Theorem* 8.1 *are not equivalent to the nondesarguesian spreads in Theorem* 3.1.

(ii) *The expanded fourth cousins in Theorem* 8.1 *are not equivalent to the expanded second or third cousins.*

Proof. (i) Assume that one of the spreads Σ^* in Theorem 3.1 is equivalent to one of those in Theorem 8.1. Then $H = \Gamma O^+(6e+2, q)_{\Sigma^*}$ has a subgroup G_N or G_M as in Theorem 3.1, as well as a subgroup with an orbit of length $|\Sigma^*|-1$ or $|\Sigma^*|-2$. Thus, H is at least 2-transitive on Σ^*. This contradicts Lemma 3.3.

(ii) Once again this follows from Lemma 3.3. □

An explicit description of expanded third cousins is given in (9.10).

9. New Kerdock sets. In [STK, § 10], new Kerdock sets were shown to exist. Similarly, by [STK, § 5], the spreads in §§ 3 and 8 also yield new Kerdock sets over any field of characteristic 2, involving matrices of an arbitrarily large size. In this section we will provide explicit examples, using expanded third cousins of desarguesian planes. Instead of starting from a spread, we will begin with a direct construction, later verifying that it arises from such a cousin.

Let $F = GF(q^{em})$, $K = GF(q^e)$ and $K' = GF(q)$, where q is even, em is odd, and $e, m \neq 1$. Let $T: F \to K$ and $T': F \to K'$ be the trace maps.

LEMMA 9.1. *If $z \in F$ and $k \in K$ then*

(i) $T'(T(z)) = T'(z)$ *and*

(ii) $T'(kzT(z)) = T'(kz^2)$.

Proof. (i) $L(z) = T'(T(z)) - T'(z)$ defines a K'-linear map $F \to K'$ such that $L(1) = 0$ and $L(z^q) = L(z)$. Then $L[\sum_{i=1}^{em} x^{q^i}) = emL(x) = L(x)$, while $\sum_{i=1}^{em} x^{q^i} \in K'$, so $L(x) = 0$.

(ii) By (i), $T'(kzT(z)) = T'(T(kzT(z))) = T'(kT(z)T(z)) = T'(kT(z)^2) = T'(T(k^{1/2}z))^2 = T'(k^{1/2}z)^2 = T'(kz^2)$, as required. □

Next, form the K'-space $F \oplus K'$. This has a natural inner product defined by $(\alpha, a) \cdot (\beta, b) = T'(\alpha\beta) + ab$. This is a nonsingular symmetric bilinear form, and admits an orthonormal basis. Fix any such basis, and use it to identify matrices and linear transformations.

Now fix $k \in K - GF(2)$, and set $k^* = 1 + k^{1/2}$.

For $s \in F$, define M_s by

$$(\alpha, a)M_s = (s^2\alpha + ksT(s\alpha) + k^*sT'(k^*s\alpha) + ak^*s,\ T'(k^*s\alpha)).$$

THEOREM 9.2. *$\{M_s | s \in F\}$ is a Kerdock set of $(em+1)\times(em+1)$ skew-symmetric matrices.*

Proof. Since

$$\begin{aligned}(\alpha, a)M_s \cdot (\alpha, a) &= T'(\alpha[s^2\alpha + ksT(s\alpha) + k^*sT'(k^*s\alpha) + ak^*s]) + aT'(k^*s\alpha)\\ &= T'(\alpha^2 s^2) + T'(\alpha ksT(s\alpha)) + T'(\alpha k^* s)T'(k^*s\alpha)\\ &\qquad + T'(\alpha ak^*s) + aT'(k^*s\alpha)\\ &= T'(\alpha^2 s^2(1+k^{*2})) + T'(k\alpha sT(\alpha s)) = 0\end{aligned}$$

by Lemma 9.1(ii) (with $z = \alpha s$), each M_s is skew-symmetric.

Assume that

$$(\alpha, a)(M_r + M_s) = 0 \tag{9.3}$$

with $r \neq s$. Then

$$T'(k^*r\alpha) = T'(k^*s\alpha) \tag{9.4}$$

and

$$r^2\alpha + krT(r\alpha) + k^*rT(k^*r\alpha) + ak^*r = s^2\alpha + ksT(s\alpha) + k^*sT(k^*s\alpha) + ak^*s. \tag{9.5}$$

Multiply (9.5) by α, and set $x = r\alpha$ and $y = s\alpha$:

$$x^2 + kxT(x) + k^*xT'(k^*x) + ak^*x = y^2 + kyT(y) + k^*yT'(k^*y) + ak^*y. \tag{9.6}$$

Apply T:

$$\begin{aligned}T(x)^2 + kT(x)^2 + k^*T(x)T'(k^*x) + T(ak^*x)\\ = T(y^2) + kT(y)^2 + k^*T(y)T'(k^*y) + T(ak^*y).\end{aligned}$$

By (9.4), $T'(k^*x) = T'(k^*y)$, so this reduces to

$$(T(x)+T(y))^2(1+k) = k^*(T(x)+T(y))T'(k^*x).$$

Now $T(x)+T(y)$ is 0 or $T'(k^*x)/k^*$. If $k^*(T(x)+T(y)) = T'(k^*x)$, apply T':

$$\begin{aligned}T'(k^*x) = T'(T(k^*T(x) + k^*T(y))) &= T'(T(k^*x + k^*y))\\ &= T'(k^*x) + T'(k^*y) = 0\end{aligned}$$

by Lemma 9.1(i) and (9.4). Thus, $T(x)+T(y) = T'(k^*x)/k^* = 0$.

This leaves us with the case $T(x)+T(y) = 0$. By (9.6) and (9.4),

$$(x+y)^2 + k(x+y)T(x) + k^*(x+y)T'(k^*x) + ak^*(x+y) = 0.$$

If $\alpha = 0$ then $a = 0$ by (9.5). Assume that $a \neq 0$, so $\alpha \neq 0$ and $x + y \neq 0$. Then

$$x + y + kT(x) + k^*T'(k^*x) + ak^* = 0.$$

Consequently, $x + y \in K$, so that $x + y = T(x) + T(y) = 0$, which is not the case. This contradiction completes the proof of Theorem 9.2. □

Remark 9.7. Let A_s be the matrix defined by $(\alpha, a)A_s = (s\alpha, a)$. Then $M_s = A_sM_1A_s$. Also, A_s is a symmetric matrix, since $(\alpha, a)A_s \cdot (\beta, b) = T(\alpha s\beta) + ab = (\alpha, a) \cdot (\beta, b)A_s$. Clearly, the matrices $\{A_s | s \in F^*\}$ form a cyclic automorphism group of the Kerdock set in Theorem 9.2 which is transitive on the nonzero members. Of course, Kerdock sets need not have such a cyclic automorphism group: none of the ones arising from Theorem 3.1 does.

Remark 9.8. In the same notation, the usual Kerdock set on $F \oplus K'$ consists of the matrices N_s, $s \in F$, defined by

$$(\alpha, a)N_s = (s^2\alpha + sT'(s\alpha) + as,\ T'(s\alpha)).$$

Once again, $N_s = A_s N_1 A_s$. The corresponding spread is the desarguesian one in (4.2).

Remark 9.9. The Kerdock set corresponding to the spread in [STK, § 8] can be described in a similar manner. Let $F = GF(q^3)$, $K = GF(q)$, and define $(\alpha, a) \cdot (\beta, b)$ as usual. This time,

$$(\alpha, a)N_s = (as^{q+q^2} + \alpha^q s^{q^2} + \alpha^{q^2} s^q,\ T(\alpha s^{q+q^2})).$$

It is an amusing exercise to verify directly that this does, indeed, yield a Kerdock set.

We now turn to the spread from which the Kerdock set of Theorem 9.2 arises.

Let F, K, K', T, T', k and k^* be as before. Let V_0 denote the F-space with basis e, f, regard V_0 as a K'-space, and form the $(2em+2)$-dimensional vector space $V' = V_0 \oplus \langle u', w' \rangle$. Define $Q' : V' \to K'$ by

$$Q'(\alpha e + \beta f + cu' + dw') = T'(\alpha\beta) + c^2 + cd.$$

This yields an $\Omega^+(2em+2, q)$ space. Set

$$\begin{aligned} &\Sigma^*[\infty] = Ff + K'(u' + w'), \\ &\Sigma^*[s] = \{\alpha e + (s^2\alpha + ksT(s\alpha) + k^*sa)f + T'(k^*s\alpha)u' + aw' \mid \alpha \in F,\ a \in K\} \end{aligned} \tag{9.10}$$

and $\Sigma^* = \{\Sigma^*[s] \mid s \in F \cup \{\infty\}\}$.

THEOREM 9.11. (i) Σ^* *is an expanded third cousin of* $AG(2, (q^e)^m)$.

(ii) *The Kerdock set defined by the pair* $(\Sigma^*, \Sigma^*[\infty])$ *is the Kerdock set in Theorem* 9.2.

(iii) *The Kerdock set in Theorem* 9.2 *is not equivalent to the desarguesian one in Remark* 9.8.

Proof. $\Sigma^* - \{\Sigma^*[\infty]\}$ can be obtained as follows. Identify $\Sigma^*[0]$ with $F \oplus K'$ in the natural manner. Let $\pi : \Sigma^*[0] \to \Sigma^*[\infty]$ be defined by $(\alpha e + aw')\pi = \alpha f + a(u' + w')$. If $s \in F$ let M_s be as in (9.2). Then

$$\Sigma^*[s] = \{\alpha e + aw' + (\alpha e + aw')M_s\pi \mid \alpha \in F,\ a \in K'\}. \tag{9.12}$$

Since M_s is skew-symmetric, $(\alpha e + aw, (\alpha e + aw')M_s\pi) = (\alpha e + aw)M_s \cdot (\alpha e + aw) = 0$. Thus, $\Sigma^*[s]$ is a totally singular $(em+1)$-space.

In order to compute $\langle u' \rangle^\perp \cap \Sigma^* / \langle u' \rangle$, set $a = 0$ and $u = 0$ in (9.10) and obtain (4.5) (with K' replacing K in (4.5)). By definition (§ 2 and [STK, § 3]), Σ^* is obtained as required in (i). Then (ii) also follows by definition [STK, § 5]. Finally, (iii) is an immediate consequence of Theorem 8.1 and [STK, Lemma 5.4]. □

THEOREM 9.13. *Let* q *be a power of* 2, *and let* $2n-1$ *be composite.*

(i) *There are at least two inequivalent nondesarguesian spreads in an* $\Omega^+(4n, q)$ *space.*

(ii) *There are at least three inequivalent nondesarguesian Kerdock sets of* $2n \times 2n$ *matrices over* $GF(q)$.

Proof. (i) Write $2n-1 = em$ with $e > 1$ and $m > 1$. If $q^{em} \neq 2^9$ we can apply Theorem 8.1 (for a suitable choice of e and m). If $q^{em} = 2^9$, use Theorem 3.1.

(ii) If Σ is one of the spreads in Theorem 9.11, then $\Gamma O^+(4n, q)_\Sigma$ is not transitive on Σ. Consequently, the result follows from [STK, Lemma 5.4]. □

10. Concluding remarks. The reader will have noticed that we have left at least as many questions unanswered as we have answered. Here is a sample of some of these questions.

(1) Prove that all of the spreads in Theorem 8.1 are inequivalent. This will require a much more geometric approach to inequivalence questions.

(2) The expansion process can be repeated indefinitely. Do new spreads always arise? Prove that they do in the case of cousins of desarguesian spreads.

In particular, the fourth cousins of desarguesian planes can be expanded and sliced over and over in such a way that each resulting translation plane has a collineation transitively permuting the points at infinity. Presumably, this produces enormous numbers of flag-transitive translation planes. (Of course, we already know at least $q/2 \log_2 q$ flag-transitive planes of order q^{2n-1} [STK, Thm. 4.2iv]. In particular, if $2n-1$ is composite, then there are more than $2^{\sqrt{n}}/2\sqrt{n}$ flag-transitive planes of order 2^{2n-1}.)

Similarly, third cousins of desarguesian planes can be expanded over and over while retaining the existence of a collineation behaving as in Proposition 6.6. (However, there are already known to be more than $2^{\sqrt{n}}/2\sqrt{n}$ planes of order 2^{2n-1} behaving this way whenever $2n-1$ is composite.)

(3) The orthogonal spreads in Theorem 8.1 do not have transitive groups, and hence produce large numbers of nonisomorphic translation planes [STK, (3.6)]. Do any of them have interesting properties? Some have the rather perverse property that no collineation acts nontrivially on the line at infinity; when expanded, these undoubtedly produce large numbers of inequivalent Kerdock sets.

(4) If Σ is as in Theorem 8.1, and if W is an $\Omega^-(2em, q)$ subspace, then $W \cap \Sigma$ is a spread of W. Show that the resulting spreads are not equivalent to spreads obtained from desarguesian $\Omega^+(2em+2, q)$ spreads.

(5) Find an internal criterion for a translation plane to be symplectic.

REFERENCES

[1] A. M. COHEN AND H. A. WILBRINK, *The stabilizer of Dye's spread on a hyperbolic quadric in* $PG(4n-1, 2)$ *within the orthogonal group*, to appear.

[2] F. DECLERCK, R. H. DYE AND J. A. THAS, *An infinite class of partial geometries associated with the hyperbolic quadric in* $PG(4n-1, 2)$, Europ. J. Combinatorics, 1 (1980), pp. 323–326.

[3] P. DEMBOWSKI, *Finite Geometries*, Springer, Berlin-Göttingen-Heidelberg, 1968.

[4] P. FONG AND G. M. SEITZ, *Groups with a* (B, N)*-pair of rank 2.I*, Invent. Math., 21 (1973), pp. 1–57.

[5] D. F. HOLT, *Transitive permutation groups in which an involution central in a Sylow 2-subgroup fixes a unique point*, Proc. LMS (3), 37 (1978), pp. 165–192.

[6] W. M. KANTOR, *Spreads, translation planes and Kerdock sets.* I, this Journal, 3 (1982), pp. 151–165.

[7] H. LÜNEBURG, *Translation Planes*, Springer, New York, 1980.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 3, September 1982

0196-5212/82/0303-0007 $01.00/0

A COMBINATORIAL PROOF OF THE ALL MINORS MATRIX TREE THEOREM*

SETH CHAIKEN†

Abstract. Let (A_{ij}), $i, j \in V$ be the matrix with entries $-a_{ij}$ if $i \neq j$ and diagonal entries such that all the column sums are zero. Let a_{ij} be a variable associated with arc ij in the complete digraph G on vertices V. Let $A(\bar{W} \mid \bar{U})$ be the matrix that results from deleting sets of k rows W and columns U from A. The all minors matrix tree theorem states that $|A(\bar{W} \mid \bar{U})|$ enumerates the forests in G that have (a) k trees, (b) each tree contains exactly one vertex in U and exactly one vertex in W, and (c) each arc is directed away from the vertex in U of the tree containing the arc. We give an elementary combinatorial proof in which we show that each of the terms in $|A(\bar{W} \mid \bar{U})|$ that corresponds to an enumerated forest occurs just once and the other terms cancel. The sign of each term is determined by the parity of the linking from U to W contained in the forest, and is easy to calculate explicitly in the proof.

The results are extended to signed graphs. The theorem provides a coordinatization (linear representation) of gammoids that is in a certain sense natural.

1. Introduction. This paper describes an elementary, combinatorial proof of the matrix tree theorem, an extension of it to signed and voltage graphs, and its applicability to the coordinatization of gammoids. We begin with a statement of the theorem.

Let the variables a_{ij}, for $i, j \in S$ and $i \neq j$ be weights on the arcs ij of the complete, loopless directed graph on a finite set of vertices S. Define matrix A by

$$A_{ij} = \begin{cases} -a_{ij} & \text{if } i \neq j, \\ \sum_k a_{kj} & \text{if } i = j. \end{cases} \tag{1}$$

A can be regarded as a "special" weighted adjacency matrix in which the jth diagonal entry is the sum of the weights of arcs directed into vertex j. Let $A(\bar{W} \mid \bar{U})$ be the submatrix of A obtained by deleting the rows indexed by the elements of $W \subset S$ and the columns indexed by $U \subset S$. Assume S is linearly ordered; for example, it may be $\{1, 2, \cdots, N\}$. Assume $|W| = |U|$. When F is a set of arcs, a_F denotes the product of their weights.

(ALL MINORS) MATRIX TREE THEOREM.

$$\det A(\bar{W} \mid \bar{U}) = \varepsilon(W, S)\varepsilon(U, S) \sum_F \varepsilon(\pi^*) a_F, \tag{2}$$

where the $\varepsilon(\cdot)$ denote signs which are defined in § 2. The sum is over all forests F such that

(i) *F contains exactly $|W| = |U|$ trees.*
(ii) *Each tree in F contains exactly one vertex in U and exactly one vertex in W.*
(iii) *Each arc in F is directed away from the vertex in U of the tree containing that arc.*

F defines a bijection or matching $\pi^*: W \to U$ so $\pi^*(j) = i$ if and only if i and j are in the same tree of F.

The all minors matrix tree theorem was given in a form similar to that here by W. K. Chen [4]. The rooted, directed forests enumerated in this theorem are sometimes called branchings, the components of which are called arborescences.

One should observe that every forest enumerated by (2) contains a collection of $|U|$ disjoint, simple, directed paths each of which starts at a vertex $i \in U$ and ends at

* Received by the editors July 15, 1981, and in revised form November 4, 1981.
† Department of Computer Science, State University of New York at Albany, Albany, New York 12222.

a vertex $\pi^{*-1}(i) \in W$. Each element of $U \cap W$ comprises a trivial path of one vertex. π^*, and therefore the relative signs of the terms in (2) are completely determined by the pairs defined by the start and end vertices of these paths.

When $U = W$ every path above degenerates to a single vertex. Every sign in (2) becomes +1. If we replace the a_{ij} by 0s or 1s, the theorem gives us a way to count the forests rooted and directed away from the vertices U in an arbitrary directed graph. The resulting theorem is an easy generalization of the classical directed graph version of the matrix tree theorem, for which $|U| = 1$. The latter was probably first described by Sylvester [23], [17], and was proved by Borchardt [2] and Tutte [24]. The undirected graph version is a special case for which $a_{ij} = a_{ji}$. When a_{ij} is given the value of the electrical conductance of the resistor joining nodes i and j in an electrical network, (2) for $|U| = 1$ and $|U| = 2$ can be used to solve the electrical network equations. The use of the duals of these "tree sums" for this purpose was given by Kirchhoff [9]. Maxwell [14, Ch. 6 and appendix] described this application of (2) which is called Maxwell's rule. See [16] for an historical survey and applications. The application of the matrix tree theorem and similar theorems to electrical network theory is detailed by Chen [4]. The interested reader should also see [13] and [22].

Let G be a directed graph with vertices S. A *linking* in G from $U \subset S$ onto $W \subset S$ is a subgraph of G consisting of $|U|$ disjoint, directed paths each of which starts at a vertex in U and ends at a vertex in W. If the a_{ij} are set to appropriate values derived from a simple modification of G, a matrix $M(S|S)$ is obtained for which $M(\bar{W}|\bar{U})$ is nonsingular if and only if there is a linking from U onto W in G. Thus submatrices of M^{-1} are coordinatizations (linear representations) of gammoids defined by G. The coordinatizations so obtained are such that (up to a $(\det M)^k$ factor, which is a polynomial with all positive terms) determinants of their minors are generating functions for directed forests that contain linkings. These generating functions have the property that the sign of each term is determined by the parity of the "permutation" defined by the linking. In § 5 this coordinatization is contrasted with two other known coordinatizations. See [25] and [21] as general references for matroid theory and linking systems.

The notion of parity as used above is made precise in § 2. In fact, our proof of the matrix tree theorem is the result of a modification and strengthening of the linkage lemma of Ingleton and Piff [8] to take parity into account, along with an application of the principle of inclusion and exclusion as used by Orlin [19] in a proof of the theorem for $U = W = \{N\}$.

It is straightforward to extend the matrix tree theorem to graphs with multiple arcs. We omit these details except in § 4 where the results are extended to signed graphs. There the results apply nontrivially even to the loops and half-arcs that may belong to such graphs.

Our proofs are purely combinatorial in that we show every expression we deal with is a generating function for a set of combinatorial objects. We classify and count, with sign, the objects that correspond to a given monomial in order to compute its coefficient. This way we can see why the subgraphs enumerated by (2) contain linkings and have no cycles. We also see that the weights of the arcs in the linking only come from off-diagonal matrix entries and all the other weights come from diagonal entries. These insights lead us to proofs of extensions of the matrix tree theorem to signed and voltage graphs ([27], [6] and [7]) which are discussed in § 4.

The author's study of the matrix tree theorem and the work in § 2 and § 5 is mostly from [3], but §§ 3 and 4 are new. [1] is a general reference for the elementary graph theory notions which we do not define explicitly.

2. Matchings, paths, cycles and signs. Let A and B be equicardinal and not necessarily disjoint subsets of a set S. All sets in this paper are finite. A bijection $\pi: A \to B$ is called a *matching*. A *k-path* in π is a sequence $(x_0, x_1, \cdots, x_k)$ for which $x_0 \in A \backslash B$, $x_k \in B \backslash A$, and $\pi(x_i) = x_{i+1}$ for $0 \leqq i < k$. A *0-path* or trivial path (x_0) in π must satisfy $x_0 \notin A \cup B$. For nontrivial k-paths, $k > 0$, the elements $x_0, x_1, \cdots, x_k$ are distinct, and $x_i \in A \cap B$ for $0 < i < k$. For $n > 0$, an *n-cycle* in π is a set of distinct x_i, $\{x_1, x_2, \cdots, x_n\}$, for which $\pi(x_i) = x_{i+1}$ for $1 \leqq i < n$ and $\pi(x_n) = x_1$. Every element of an n-cycle in π belongs to $A \cap B$. A 1-cycle is called a trivial cycle.

We can view the matching π as a directed graph on S in which ij is an arc if and only if $i \in A$ and $\pi(i) = j$. Given a directed graph G, we say π is a matching in G when $\pi(i) = j$ only if $ij \in G$. Unless otherwise specified, a cycle or path will always mean a directed cycle or path. When we use the terms circuit or (connected) component, we ignore the arc directions.

It is clear that every matching decomposes into disjoint paths and cycles. (To be technical, we should note that the trivial paths depend upon the underlying set S.) The outdegree (resp. indegree) of i in π is 1 if $i \in A$ (resp. $i \in B$) and is 0 otherwise. When $A = B$ there are no nontrivial paths in π and we get the familiar decomposition of a permutation of A into cycles.

For completeness, we state the linkage lemma [8]. A *linking* of U onto W is a collection of $|U|$ disjoint directed paths each of which starts at an element of U and ends at an element of W.

LEMMA. *Suppose G is a directed graph of S. Let $G' = G \cup \{ii \mid i \in S\}$. Suppose U, $W \subset S$. Then, there is a linking in G from U onto W if and only if there is a matching $\pi: S \backslash W \to S \backslash U$ in G'.*

For a proof, see [25].

Now suppose A and B are linearly ordered; for example, suppose A and B are sets of integers. The pair $\{i, j\} \subset A$ is an *inversion* in π if $i < j$ and $\pi(i) > \pi(j)$. Let $n(\pi)$ denote the number of inversions in π. We define the *sign* $\varepsilon(\pi)$ of the matching by

$$\varepsilon(\pi) = (-1)^{n(\pi)}.$$

When π is a permutation, it is well known that $\varepsilon(\pi)$ is its sign, that $\varepsilon(\pi)$ does not depend on the ordering of $A = B$, and that when π is decomposed into cycles,

$$\varepsilon(\pi) = \prod_{n\text{-cycles}} (-1)^{n-1}. \tag{3}$$

Let Y be a linearly ordered set and $X \subset Y$. We define

$$n(X, Y) = |\{\{i, j\} | i < j, i \in Y \backslash X, j \in X\}|$$

and

$$\varepsilon(X, Y) = (-1)^{n(X,Y)}.$$

When $Y = \{1, 2, \cdots, N\}$, $n(X, Y)$ equals $\sum X - |X| - \binom{|X|}{2}$. Hence $\varepsilon(X, Y)\varepsilon(X', Y)$ commonly appears as $(-1)^{\Sigma X + \Sigma X'}$, when $|X| = |X'|$.

Suppose S, T are linearly ordered sets and $S \cap T = \phi$. Suppose $\pi: A \to B$ and $\pi': \bar{A} \to \bar{B}$ are matchings where $A \subset S$, $\bar{A} = S \backslash A$, $B \subset T$, and $\bar{B} = S \backslash B$. We can combine π and π' to form a matching $\pi \oplus \pi': S \to T$ for which

$$\pi \oplus \pi'(i) = \begin{cases} \pi(i) & \text{if } i \in A, \\ \pi'(i) & \text{if } i \in \bar{A}. \end{cases}$$

It is easy to prove by induction on $n(A, S)+n(B, T)$ that

$$\varepsilon(\pi \oplus \pi') = \varepsilon(A, S)\varepsilon(B, T)\varepsilon(\pi)\varepsilon(\pi').$$

COROLLARY. *Suppose S is linearly ordered, $A \subset S$, $\bar{A} = S \backslash A$, $B \subset S$, and $\bar{B} = S \backslash B$. Let $\pi : A \to B$ and $\pi' : \bar{A} \to \bar{B}$ be matchings. Then*

$$\varepsilon(\pi \oplus \pi') = \varepsilon(A, S)\varepsilon(B, S)\varepsilon(\pi)\varepsilon(\pi'). \tag{4}$$

Proof. Let T be a disjoint copy of S. Redefine B, $\bar{B}$ appropriately and apply the above remark. □

Let A and B be subsets of a linearly ordered set S and $\pi : A \to B$ be a matching. The paths in π determine a matching $\pi^* : \bar{A} \to \bar{B}$ as follows: For each (possibly trivial) path $(x_0, x_1, \cdots, x_k)$, we have $\pi^*(x_k) = x_0$. The linkage lemma asserts that there is a matching $\pi : A \to B$ in a certain digraph G' if and only if there is a linking in G from $\bar{B}$ onto $\bar{A}$ which defines π^* as shown. Our strengthening of this lemma shows how the signs of any such pair π, π^* must be related.

THEOREM. *Suppose $\pi : A \to B$ and $\pi^* : \bar{A} \to \bar{B}$ are given as above. Then*

$$\varepsilon(\pi) = \varepsilon(\pi^*)\varepsilon(A, S)\varepsilon(B, S) \prod_{\substack{k\text{-paths} \\ \text{in } \pi}} (-1)^k \prod_{\substack{n\text{-cycles} \\ \text{in } \pi}} (-1)^{n-1}. \tag{5}$$

Proof. $\pi \oplus \pi^* : S \to S$ is a permutation. Its cycles consist of one $(k+1)$-cycle for each k-path in π, along with all the n-cycles of π. Hence, when we apply (3) we obtain

$$\varepsilon(\pi \oplus \pi^*) = \prod_{\substack{k\text{-paths} \\ \text{in } \pi}} (-1)^k \prod_{n\text{-cycles in } \pi} (-1)^{n-1}.$$

The identity follows immediately from (4). □

The matrix tree theorem will be an easy consequence of the decomposition of π into paths and cycles, formula (5), and the definition of the determinant

$$\det M(A \mid B) = \sum_{\pi : A \to B} \varepsilon(\pi) \prod_{i \in A} M_{i,\pi(i)}.$$

The sum is taken over all matchings $\pi : A \to B$.

3. Proof of the matrix tree theorem. For convenience, we here restate the matrix tree theorem.

ALL MINORS MATRIX TREE THEOREM.

THEOREM. *Suppose $A(S \mid S)$ is given by* (1), *the $\varepsilon(\ )$ are defined in* § 2, *and U, $W \subset S$ with $|U| = |W|$. Then*

$$\det A(\bar{W} \mid \bar{U}) = \varepsilon(W, S)\varepsilon(U, S) \sum_{F} \varepsilon(\pi^*) a_F \tag{2}$$

where the sum is over all forests F on S such that

(i) *F contains exactly $|U| = |W|$ trees.*

(ii) *Each tree in F contains exactly one vertex in U and exactly one vertex in W.*

(iii) *Each arc in F is directed away from the vertex in U of the tree containing that arc.*

F defines a matching $\pi^* : W \to U$ so $\pi^*(j) = i$ if and only if i and j are in the same tree of F.

Proof. By definition of $\det A(\bar{W} \mid \bar{U})$,

$$\det A(\bar{W} \mid \bar{U}) = \sum_{\pi : \bar{W} \to \bar{U}} \varepsilon(\pi) \prod_{i \in \bar{W}} A_{i,\pi(i)}. \tag{6}$$

Suppose in (6), for each matching π, we distinguish the diagonal entries, which have the form A_{jj}, from the off-diagonal entries of A. If we apply the definition of A, we obtain

$$\det A(\bar{W}\,|\,\bar{U}) = \sum_{(\pi,\sigma)} \varepsilon(\pi)\Big[\prod_{ij\in\sigma} a_{ij}\Big] \prod_{\substack{\pi(i)=j \\ i\neq j}} (-a_{ij}) \quad . \tag{7}$$

Here, the determinant is expressed as a sum of terms $\pm a_H$, one for each pair (π, σ) such that π is a matching $\pi: \bar{W} \to \bar{U}$ and σ is a set of arcs consisting of one and only one arc ij for each j such that $\pi(j) = j$.

Let H be any subgraph defined by a pair (π, σ) as above. In H, for all $j \in S$,

$$\text{indeg}(j) = \begin{cases} 1 & \text{if } j \in \bar{U}, \\ 0 & \text{if } j \in U. \end{cases} \tag{8}$$

The indegrees in H are all at most one. Hence, any circuit in H must be a (directed) cycle. Furthermore, the cycles in H are disjoint. Now consider any path P in π as a subgraph of H. No arc in P can belong to a cycle in H. This is because the indegree in P of each vertex in P is equal to its indegree in H. Therefore, only arcs in P may be directed into vertices in P. We conclude that each cycle in H either belongs to σ or is a nontrivial cycle in π.

We can now conclude that if H has no cycles, then H is a forest F that satisfies (i), (ii), and (iii). Let us therefore write $\det A(\bar{W}\,|\,\bar{U})$ as $\sum_H c_H a_H$. The theorem will be proved when we show that $c_H = 0$ when H contains a cycle, that c_H is given by (2) otherwise, and that there is a pair (π, σ) that defines $H = F$ for every forest that satisfies (i), (ii), and (iii).

Let π^* be the matching $\pi^*: W \to U$ defined in (2) by the paths in π. When we apply (5) to (7) we obtain

$$\det A(\bar{W}\,|\,\bar{U}) = \varepsilon(\bar{W}, S)\varepsilon(\bar{U}, S) \sum_{(\pi,\sigma)} \varepsilon(\pi^*)(-1)^{cy(\pi)}\Big[\prod_{ij\in\sigma} a_{ij}\Big] \prod_{\substack{\pi(i)=j \\ i\neq j}} a_{ij} \tag{9}$$

where $cy(\pi)$ is the number of nontrivial cycles in π.

Let H be a subgraph with K cycles that is defined by some (π_1, σ_1). Let us consider all pairs (π, σ) that define H. In each pair, π has the same paths as π_1. All the arcs that are neither in a cycle nor in a path in π_1 belong to σ. Each cycle in H can be either a nontrivial cycle in π or a cycle in σ. Hence, there are 2^K pairs (π, σ) that define H and

$$c_H = \varepsilon(\bar{W}, S)\varepsilon(\bar{U}, S)\varepsilon(\pi^*) \sum_{c=0}^{K} (-1)^c \binom{K}{c} = \pm(1-1)^K = \begin{cases} \pm 1 & \text{if } K = 0, \\ 0 & \text{if } K \neq 0. \end{cases}$$

It is easy to see $\varepsilon(\bar{W}, S)\varepsilon(\bar{U}, S) = \varepsilon(W, S)\varepsilon(U, S)$ when $|W| = |U|$. Hence, c_H is given by (2).

Finally, suppose F is a forest that satisfies (i), (ii), and (iii). F is defined by the pair (π, σ) for which π has the paths linking U to W in F, π has no nontrivial cycles, and σ consists of all the arcs in F not in these paths. □

The last step in the proof tells us each F counted by (2) is due to just one matching π in (6). The weights of the arcs in the linking only come from the off-diagonal entries of A. All the other arc weights come from diagonal entries which correspond to trivial cycles in π.

4. Extension to signed graphs. A signed graph is a graph to which each arc has been given a sign. See [27] for a systematic treatment of the definitions, properties, and applications of signed graphs. Broadly speaking, signed graphs differ from ordinary graphs in the matroids they define. For example, a circle (i.e., a circuit in the underlying graph) is a circuit in the signed graphic matroid only if it is *positive*—that is, the product of the signs is + (see [7]), otherwise the circle is an independent set.

A signed directed graph is like an ordinary directed graph, except each arc e is given a sign $s(e) \in \{+, -\}$, and, this time, we allow multiple arcs, *loops* (arcs of the form $e = ii$), and *half-arcs* ($e = i$; the sign of a half-arc is undefined). As in an ordinary directed graph, arc $e = ij$ is said to be directed "out" from i and "into" j (even if $i = j$). If $e = i$, e is said to be directed into i. A directed k-path is a sequence of arcs $(e_1 = x_0x_1, e_2 = x_1x_2, \cdots, e_k = x_{k-1}x_k)$ in which all the x_i are distinct. A directed n-cycle is a set of n arcs $\{e_1 = x_1x_2, e_2 = x_2x_3, \cdots, e_n = x_nx_1\}$ incident on n distinct vertices. Note half-arcs cannot appear in (directed) k-paths or n-cycles, while a loop is a 1-cycle. A signed directed graph differs from a signed graph (as in [27]) in that the fixed order of the endpoints of each arc allows us to define directed paths and cycles in directed graphs. These definitions must not be confused with those involving oriented signed graphs [26].

A path or cycle will be called *positive* if the product of the signs of its arcs is +; it is negative otherwise.

In this chapter we extend the matrix tree theorem and our proof to signed directed graphs. Then, in the same way the undirected graph version of the matrix tree theorem was obtained from the directed graph version, we obtain an extension of the matrix tree theorem to signed graphs by Zaslavsky [27]. We further extend the theorem to voltage graphs [6] over an abelian group.

As for the matrix tree theorem, we assign a weight a_e to each arc in the signed directed graph. One must not confuse the weight of an arc with its sign. Matrix $A(S \mid S)$ is defined as follows.

$$\text{If } i \neq j, \qquad A_{ij} = -\sum_e s(e)a_e \tag{10a}$$

where the sum is over all arcs $e = ij$.

$$A_{jj} = \sum_e a_e + \sum_l 2a_l + \sum_h a_h \tag{10b}$$

where e ranges over arcs ij directed into j for which $i \neq j$, l ranges over negative loops jj, and h ranges over half-arcs into j.

MATRIX TREE THEOREM FOR SIGNED DIRECTED GRAPHS. *Let G be a signed directed graph on S and $A(S \mid S)$ be as above. Suppose U, $W \subset S$, $|U| = |W|$. Then*

$$\det A(\bar{W} \mid \bar{U}) = \varepsilon(U, S)\varepsilon(W, S) \sum_F \varepsilon(\pi^*)(-1)^{np(F)} 2^{nc(F)} a_F \tag{11}$$

where the sum is over all sets of arcs F in G such that

(i) *F contains $|U| = |W|$ components that are trees.*

(ii) *Each tree from* (i) *contains exactly one vertex in U and one vertex in W.*

(iii) *Each arc in each tree from* (i) *is directed away from the vertex in U of the tree containing that arc. Hence these trees together contain a linking from U onto W. This linking defines $\pi^*: W \to U$ as in the matrix tree theorem. $np(F)$ is the number of negative paths in this linking.*

(iv) *Each of the remaining components of F contains exclusively either just one half-arc or just one negative (directed) cycle. There are no other circles and each*

remaining arc is directed away from the half-arc or (directed) cycle of its component. $nc(F)$ is the number of negative cycles.

Proof. It is easy to verify that

$$\det A(\bar{W} \mid \bar{U}) = \sum_{H} c_H a_H,$$

where the sum is over some subgraphs H in which for all $j \in S$, (8) is satisfied. Since our task is to determine c_H, we can set $a_e = 0$ for $e \notin H$ and write our proof as in § 3. Please note that ij designates a particular arc in $H \subset G$ with a given sign. Equation (7) becomes ($\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ if $i \neq j$, and $\delta_i = 0$)

$$\det A(W \mid U) = \sum_{(\pi, \sigma)} \varepsilon(\pi) \left[\prod_{ij \in \sigma} (1 + \delta_{ij}) a_{ij} \right] \left[\prod_{\substack{\pi(i) = j \\ i \neq j}} (-s(ij) a_{ij}) \right] \tag{12}$$

where we have abused the notation because σ may contain a half-arc. Still, any nontrivial directed cycle in H is either a nontrivial cycle in π or a nontrivial directed cycle in σ. The arc sign factors $s(\cdot)$ only occur for arcs in π, so the extension of (9) is

$$\det A(\bar{W} \mid \bar{U}) = \varepsilon(\bar{W}, S) \varepsilon(\bar{U}, S)$$

$$\cdot \sum_{(\pi, \sigma)} \varepsilon(\pi^*) (-1)^{cy(\pi)} (-1)^{nc'(\pi)} (-1)^{np(\pi)} \left[\prod_{ij \in \sigma} (1 + \delta_{ij}) a_{ij} \right] \left[\prod_{\substack{\pi(i) = j \\ i \neq j}} a_{ij} \right] \cdot \tag{13}$$

where $nc'(\pi)$ and $np(\pi)$ are respectively the numbers of negative nontrivial cycles and negative paths in π. If H has K_p positive nontrivial directed cycles and K_n negative nontrivial directed cycles, there are $2^{K_p + K_n}$ pairs (π, σ) that define H. For each trivial cycle jj in H, $jj \in \sigma$ and $\pi(j) = j$ for each (π, σ) that defines H, and so the factor $(1 + \delta_{jj}) = 2$ occurs in each term for H in (13). Let K_t be the number of trivial cycles in H.

We conclude

$$c_H = \varepsilon(\bar{W}, S) \varepsilon(\bar{U}, S) \varepsilon(\pi^*) (-1)^{np(H)} 2^{K_t} (1 + 1)^{K_n} (1 - 1)^{K_p}. \tag{14}$$

Thus, if H has no positive cycles, c_H is given by (11). Finally, suppose F is given which satisfies (i), (ii), (iii), and (iv) with K_n negative nontrivial directed cycles. Again, we set all the as but those in a_F to zero. Then there are 2^{K_n} pairs (π, σ) that define F. In all of them, π contains the linking described in (iii) and σ contains the negative trivial directed cycles and all arcs neither in a cycle nor the linking. Each negative, nontrivial directed cycle belongs to either π or σ exclusively. Thus a_F appears in (11). □

For a signed (undirected) graph G on S, $A(S \mid S)$ is a symmetric matrix [27]. To represent G by a signed directed graph G', we represent each undirected arc $e = ij$ by a pair of directed arcs ij and ji with identical weights a_e and signs, even if $i = j$. Half arcs in the undirected graph are represented by only one arc in the directed graph. Hence the analogue of (10) is

$$\text{if } i \neq j, \qquad A_{ij} = -\sum_{e} s(e) a_e, \qquad A_{jj} = \sum_{e} a_e + \sum_{l} 4 a_l + \sum_{h} a_h.$$

The factor of 4 makes more sense when A is written $A = DED^t$ where D is a signed incidence matrix of G and E is the diagonal matrix of arc weights.

MATRIX TREE THEOREM FOR SIGNED (UNDIRECTED) GRAPHS [27].

$$\det A(\bar{A} \mid \bar{U}) = \varepsilon(U, S)\varepsilon(W, S) \sum_{F} \varepsilon(\pi^*)(-1)^{np(F)} 4^{nc(F)} a_F. \tag{15}$$

The sum is over all sets of arcs F that satisfy conditions similar to (i), (ii), (iii) *and* (iv). *The new conditions are obtained by deleting the "directed" qualifier everywhere from the old conditions.*

Proof. Suppose we apply the directed graph version of the theorem to G'. Suppose T is a tree in G that, according to the conditions, contains $u \in U$ or a half arc. Then there is exactly one directed tree T' in G', with $a_T = a_{T'}$, that satisfies the corresponding conditions, and conversely. Now suppose T is a subgraph in G that, according to condition (iv), contains a unique circle (which is negative). Then there are just two subgraphs T' in G', with $a_T = a_{T'}$ that satisfy the corresponding conditions, and conversely. The directed cycles in these two subgraphs are directed oppositely while all the other directed arcs are identical. Thus, each undirected graph F that satisfies (i), (ii), (iii) and (iv) with $nc(F)$ negative circles is counted $2^{nc(F)}$ times by directed graphs F' in G' with $a_F = a_{F'}$. The coefficient for each directed graph F' is $\pm 2^{nc(F)}$ (and is constant), so c in (15) is $\pm 2^{nc(F)} 2^{nc(F)} = \pm 4^{nc(F)}$. □

A voltage graph ([27], [6]) is a graph to which each arc has been given an element of a group. Signed graphs are a special case of voltage graphs. Our method can be used to prove a version of the matrix tree theorem for voltage graphs over an abelian group Γ. It is necessary to extend the ring of coefficients for the polynomials in the arc weights to the group ring of Γ. A directed cycle is positive when the product of the voltages on its arcs is 1, the identity of Γ. Suppose we define matrix A for a voltage graph as in (10) except $s(e)$ now stands for the voltage of arc e and the coefficient of a_l in A_{jj} when l is a loop $l = jj$ is $(1 - s(l))$. When E is a set of arcs, let $s(E)$ denote the product of their voltages. The voltage directed graph version goes through as for the signed directed graph theorem except that the notion of positivity is replaced with the notion of positivity for voltage graphs and expression (11) becomes

$$\det A(\bar{W} \mid \bar{U}) = \varepsilon(U, S)\varepsilon(W, S) \sum_{F} \varepsilon(\pi^*) s(P) \prod_{C} (1 - s(C)) a_F.$$

Here, P is the linking from U onto W in condition (iii). C ranges over the nonpositive directed cycles in F.

5. Gammoids. The matrix tree theorem can be used to give a coordinatization (i.e., representation of a matroid by the column vectors of a matrix over a field) of gammoids that is "natural" with respect to sign in a way that other known coordinatizations are not. We discuss this below. The books by Welsh [25] and Schrijver [21] are our references for matroids and linking systems.

Let G be a directed graph on vertices S and let a_{ij} be an indeterminate when ij is an arc in G and be zero otherwise. The matrix tree theorem implies that $A(\bar{W} \mid \bar{U})$ is nonsingular only if there is a linking in G of U onto W.

Now let $-B$ be the same matrix as A except that its main diagonal entries are all zero. Let I be the identity matrix and $T = I - B$. The linkage lemma of Ingleton and Piff [8] asserts that $\det T(\bar{W} \mid \bar{U})$ is nonzero if and only if there is a linking in G of U onto W. The subsets U of S for which there is a linking in from U onto W, where W is a fixed subset of S, comprise the bases of a matroid. Such a matroid is called a *strict gammoid* [20]. The linkage lemma is the key step in the proof that a matroid is a strict gammoid if and only if it is the dual of a transversal matroid.

Linking systems or bimatroids [10] provide an alternative view of matroid theory that is most suitable for the purposes of this section. A linking system (X, Y, Λ) is equivalent to a matroid M on the disjoint union $X \cup Y$ with a distinguished base X. A pair (U, W) belongs to $\Lambda \subset 2^X \times 2^Y$, which is called the set of linked pairs (or nonsingular minors), when $(X \backslash U) \cup W$ is a base in M. The axioms for linking systems given by Schrijver [21] are properties satisfied by the (U, W) such that there is a matching from U onto W in a bipartite graph $G \subset X \times Y$.

(a) If $(U, W) \in \Lambda$ and $x \in U$, then $(U \backslash x, W \backslash y) \in \Lambda$ for some $y \in W$.

(b) If $(U, W) \in \Lambda$ and $y \in W$, then $(U \backslash y, W \backslash y) \in \Lambda$ for some $x \in U$.

(c) If $(U_1, W_1), (U_2, W_2) \in \Lambda$, then there exists $(U', W') \in \Lambda$ with $U_1 \subset U' \subset U_1 \cup U_2$ and $W_2 \subset W' \subset W_1 \cup W_2$.

The third is the Dulmage–Mendelsohn [15] property.

A linking system (X, Y, Λ) is said to be *coordinatized* by a matrix $M(X \mid Y)$ when $(U, W) \in \Lambda$ if and only if $M(U \mid W)$ is nonsingular. Now suppose (X, Y, Λ) is such that $(X, Y) \in \Lambda$. Schrijver shows then that (Y, X, Λ^{-1}) is a linking system, where

$$\Lambda^{-1} = \{(W, U) \mid (X \backslash U, Y \backslash W) \in \Lambda\}.$$

(Y, X, Λ^{-1}) is called the inverse of (X, Y, Λ). It follows from Jacobi's theorem [18] that if $M(X \mid Y)$ coordinatizes (X, Y, Λ), then $M^{-1}(Y \mid X)$ coordinatizes (Y, X, Λ^{-1}). To be more specific in our application of Jacobi's theorem, if $M(S \mid S)$ is any matrix and $\hat{M}(S \mid S)$ is defined by

$$\hat{M}_{ij} = \varepsilon(i, S)\varepsilon(j, S) \det M(\bar{j} \mid \bar{i})$$

(note $\varepsilon(i, S)\varepsilon(j, S) = (-1)^{i+j}$ when $S = \{1, 2, \cdots, N\}$), then

$$\det \hat{M}(U \mid W) = (\det M)^{|U|-1} \varepsilon(U, S)\varepsilon(W, S) \det M(\bar{W} \mid \bar{U}). \tag{16}$$

Let G be a directed graph on S. G defines the *strict gammoid linking system* (S, S, Λ) in which $(U, W) \in \Lambda$ if and only if there is a linking of U onto W in G. Thus, the transposed submatrices of a coordinatization of the strict gammoid linking system (S, S, Λ) comprise coordinatizations of all the gammoid matroids that can be defined by G. We will give three coordinatizations of the strict gammoid linking system defined by G. The coordinatizations will be over any extension field that contains the algebraically independent elements $\{a_e \mid e \text{ is an arc in } G\}$.

The first coordinatization is $\hat{T}$. Essentially, it was described by Schrijver and the proof of its correctness uses the linkage lemma. When we combine (16) with an argument similar to that in § 3, we obtain

$$\det \hat{T}(U \mid W) = (\det T)^{|U|-1} \sum_F \varepsilon(\pi^*)(-1)^{cy(F)} a_F$$

where the sum is over all subgraphs F of G whose connected components consist of a linking from U onto W, isolated vertices, and $cy(F)$ disjoint (directed) cycles. The linking defines a matching $\pi^*: W \to U$ where $\pi^*(j) = i$ when the linking contains a path from i to j.

The second coordinatization is $\hat{H}$, where $H = I + A$ and A is the matrix (1) in the matrix tree theorem.

THEOREM. *$H(\bar{W} \mid \bar{U})$ is nonsingular if and only if there is a disjoint collection of directed paths linking U onto W in G.*

Proof. Let $0 \notin S$. Consider graph G' on $S \cup \{0\}$ which contains all the arcs in G along with all arcs $0j$, $j \in S$. Suppose the latter arcs have weight 1. H is the submatrix of the "special" adjacency matrix (1) of G' obtained by deleting row and column 0.

If there is a linking L in G from U onto W, then there is a term in $\det H(\bar{W} \mid \bar{U})$ corresponding to the forest consisting of the arcs in L along with all arcs $0j$ for j not a vertex in L. Conversely, if $\det H(\bar{W} \mid \bar{U}) \neq 0$ there is a forest in G' that contains a linking in G from U onto W. □

The above proof along with the matrix tree theorem and (16) is used to derive

$$\det \hat{H}(U \mid W) = (\det H)^{|U|-1} \sum_{F} \varepsilon(\pi^*) a_F.$$

Apart from the $(\det H)^{|U|-1}$ factor, this is the generating function for all directed forests in G that contain linkings from U to W. The sign of each term is the sign of the matching $\pi^*: W \to U$ that the corresponding linking determines. In this sense we remark that the coordinatization $\hat{H}$ is "natural" in a way the first coordinatization fails to be.

The third coordinatization comes from Mason [12]. It is the matrix $P(S \mid S)$ defined by

$$P_{ij} = \sum_{P} a_P$$

where the sum is over all (simple) directed paths from i to j in G. Suppose $|U| = |W| = l$. Mason's proof uses Menger's theorem to factor $P(U \mid W)$ into a product of an $l \times k$ and a $k \times l$ matrix with $k < l$ when no linking from U to W exists. Lindström [11] attempted to give a proof based upon the claim that $\det P(U \mid W)$ was equal to

$$\sum_{L} \varepsilon(\pi^*) a_L \tag{17}$$

where the sum is over all linkings from U onto W and $\pi^*: W \to U$ is the matching defined by each. This claim is false when G contains directed cycles. For example, suppose G is itself a directed n-cycle. Then $S = \{1, 2, \cdots, n\}$ and the arcs of G are $\{ij \mid 1 \leqq i \leqq n \text{ and } j = i+1 \bmod n\}$, so

$$P_{ij} = \begin{cases} a_{i,i+1} a_{i+1,i+2} \cdots a_{j-1,j} & \text{if } i \neq j, \\ 1 & \text{if } i = j \end{cases}$$

where the subscripts are taken mod n. We have

$$\det P = (1 - a_G)^{n-1}. \tag{18}$$

We remark that the determinant of a submatrix of P for an acyclic graph has been applied to an enumeration problem for plane partitions by Gessel [5]. There, the relevant $\varepsilon(\pi^*)$ are all equal to 1.

It is tempting to ask whether the coordinatization $\hat{M} = \hat{T}$, $\hat{M} = \hat{H}$ or P can be "fixed up" so that the factor $(\det M)^{|U|-1}$ no longer appears in $\det \hat{M}(U \mid W)$ in the former two or that (17) indeed is the determinant of the $(U \mid W)$ minor in the latter. We remark the answer is no in all cases. The reason is simply that if we require this of the 1×1 minors of the coordinatizations, we obtain the same matrices $\hat{T}$, $\hat{H}$ and P. One can ask, however, for a nice combinatorial description of $\det P(U \mid W)$ for all U, $W \subset S$, $|U| = |W|$, which will provide a combinatorial proof of (18). This question is apparently open.

REFERENCES

[1] C. BERGE, *Graphs and Hypergraphs*, North-Holland, Amsterdam, 1973.

[2] C. W. BORCHARDT, *Ueber eine der Interpolation entsprechende Darstellung der Eliminations-Resultante*, J. Reine Angew. Math., 57 (1860), pp. 111–121.

[3] S. CHAIKEN, *Matrix free theorems and degree sequence realization by strongly 2-connected digraphs*, Ph.D. thesis, Massachusetts Institute of Technology, 1980.

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

[5] I. M. GESSEL, *Determinants and plane partitions*, preprint.

[6] J. L. GROSS, *Voltage graphs*, Discrete Math., 9 (1974), pp. 239–246.

[7] F. HARARY, *On the notion of balance of a signed graph*, Michigan Math. J., 2 (1953–1954), pp. 143–146.

[8] A. W. INGLETON AND M. J. PIFF, *Gammoids and transversal matroids*, J. Combin. Theory Ser. B, 15 (1973), pp. 51–68.

[9] G. KIRCHHOFF, *Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Verteilung galvanischer Ströme gefuhrt wird*, Ann. Physik Chemie, 72 (1847), pp. 497–508; *On the solution of the equations obtained from the investigation of the linear distribution of Galvanic currents* (J. B. O'Toole, tr.) IRE Trans. Circuit Theory, 5 (1958), pp. 4–8.

[10] J. P. S. KUNG, *Bimatroids and invariants*, Adv. in Math., 30 (1978), pp. 238–249.

[11] B. LINDSTRÖM, *On the vector representations of induced matroids*, Bull. London Math. Soc., 5 (1973), pp. 85–90.

[12] J. H. MASON, *On a class of matroids arising from paths in graphs*, Proc. London Math. Soc., (3), 25 (1972), pp. 55–74.

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

[14] J. C. MAXWELL, *Electricity and Magnetism*, Clarendon Press, Oxford, 1892. Reprinted by Dover, New York.

[15] N. S. MENDELSOHN AND A. L. DULMAGE, *Some generalizations of the problem of distinct representatives*, Canad. J. Math., 10 (1958), pp. 230–241.

[16] J. W. MOON, *Counting labelled trees*, Canadian Math. Monograph 1, Canadian Math. Congress, 1970.

[17] T. MUIR, *Theory of Determinants in the Historical Order of Development*, vol. 2, Macmillan, London, 1911.

[18] ———, *A Treatise on the Theory of Determinants*, Dover, New York, 1960.

[19] J. B. ORLIN, *Line-digraphs arborescences, and theorems of Tutte and Knuth*, J. Combin. Theory Ser. B, 25 (1978), pp. 187–198.

[20] H. PERFECT, *Application of Menger's theorem*, J. Math. Anal. Appl., 22 (1968), pp. 96–111.

[21] A. SCHRIJVER, *Matroids and linking systems*, Math. Centre Tract 88, Mathematical Centre, Amsterdam, 1978.

[22] C. A. B. SMITH, *Electric currents in regular matroids*, in Combinatorics, D. J. A. Welsh and D. R. Woodal, eds., Inst. of Math. and Appl., Southend-on-Sea, 1972, pp. 262–284.

[23] J. J. SYLVESTER, *On the change of systems of independent variables*, Quart. J. Pure Appl. Math., 1 (1855), pp. 42–56. Also appears in Collected Math. Papers, Cambridge, 2 (1908), pp. 65–85, and is reviewed in [17].

[24] W. T. TUTTE, *The dissection of equilateral triangles into equilateral triangles*, Proc. Cambridge Philos. Soc., 44 (1948), pp. 463–482.

[25] D. J. A. WELSH, *Matroid Theory*, Academic Press, New York, 1976.

[26] T. ZASLAVSKY, *Orientation of signed graphs*, preprint.

[27] ———, *Signed graphs*, Discrete Appl. Math., 4 (1982), no. 1, to appear.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 3, September 1982

0196-5212/82/0303-0008 $01.00/0

A CLASS OF PERFECT GRAPHS ASSOCIATED WITH PLANAR RECTILINEAR REGIONS*

MICHAEL SAKS†

Abstract. A class of graphs, arising in connection with a covering problem for rectilinear regions, is shown to be perfect. This affirms a conjecture of Chaiken, Kleitman, Saks and Shearer [SIAM J. Alg. Disc. Meth., 2 (1981), pp 394–410].

1. Introduction. Let S denote the set of unit squares in the plane whose centers have integer coordinates and whose sides are parallel to the coordinate axes. Squares are referred to by the coordinates of their centers. A finite subset T of S is called a *region*; a rectangular shaped subset of a region T is called a *rectangle of* T. Associated with a region T is a graph $G = G^T$ on vertex set T with two vertices joined by an arc if they are contained in a common rectangle of T. It is not difficult to show that the maximal cliques in this graph are exactly the maximal rectangles of T. Stable sets in G are called *antirectangles*.

In response to a question of Chvátal, Chung [cited in 2] exhibited a simply connected region T for which $\theta(G)$, the clique cover number of G (which is the minimum number of rectangles whose union is T), is strictly greater than $\alpha(G)$, the stability number of G (which is the size of the largest antirectangle of T). Chaiken et al. [2] proved that if T is *convex*, in the sense that each horizontal or vertical line of T consists of consecutive squares, then $\theta(G) = \alpha(G)$. In their paper, they noted that for convex T, the graph G need not be perfect, i.e., there may be subsets $F \subseteq T$ for which the induced subgraph G_F satisfies $\theta(G_F) > \alpha(G_F)$. They did show, however, that for the set C of *corner squares* of a convex region T (those with at least two neighboring squares not in T), the vertex induced subgraph G_C is perfect. They conjectured that for B the set of boundary squares (those with at least one neighboring square not in T), G_B is perfect. In this paper, this conjecture is settled in the affirmative.

We define a *board* to be a pair (T, F) where T is a region and $F \subseteq T$. A board is depicted by placing a blackened square in the center of each square belonging to F. A board (T, F) is said to be *convex* if T is convex. A *rectangle cover of* (T, F) is a set of rectangles of T whose union contains F and an *antirectangle of* (T, F) is an antirectangle of T contained in F. It is easy to see that $\theta(G_F)$ equals the size of the minimum rectangle cover of (T, F) and $\alpha(G_F)$ is the size of the largest antirectangle of (T, F). The main theorem of this paper is

THEOREM 1.1. *Let T be a convex region with boundary squares B. For $F \subseteq B$, the minimum rectangle cover of (T, F) has the same size as the maximum antirectangle of (T, F). Thus the subgraph G_B^T induced on G^T by B is perfect.*

In [2], it was pointed out that the graph G_B need not be perfect if T is not convex.

2. Definitions and simple facts. Let T be a convex subset of S. We identify T with the region it maps out in the plane, thus T has a *boundary* and *corners*. A *boundary segment* is a maximal line segment belonging to the boundary. A *boundary square* of T is a square which has at least one side on a boundary segment and a

* Received by the editors August 26, 1981, and in revised form October 15, 1981. The results of this paper are contained in the author's doctoral thesis completed at the Massachusetts Institute of Technology under the direction of Daniel J. Kleitman.

† University of California, Los Angeles, California 90024. Current address: Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903.

corner square is one with at least two sides on boundary segments. Corner squares are of four *types*: lower left, upper left, lower right and upper right (a corner square lying on three boundary segments has two types).

We will assume that T is *connected*, i.e., for every pair of squares s and t there is a sequence of squares $s = s_1, s_2, \cdots, s_k = t$ such that s_i and s_{i+1} share a side. It is not difficult to see that if Theorem 1.1 holds for connected regions it also holds for disconnected ones.

For $x_1 \leqq x_2 \in Z$, we use the notation $[x_1, x_2]$ to mean the set of integers between x_1 and x_2, inclusive, and $[x_1, x_2]\times[y_1, y_2]$ denotes the rectangle $\{(x, y)|x \in [x_1, x_2], y \in [y_1, y_2]\}$.

Let x^L and x^R (for left and right) denote the minimum and maximum first coordinates of any square in T and let y^D and y^U (for down and up) be the minimum and maximum second coordinates of any square in T. For $x^L \leqq x \leqq x^R$, the xth *column of* T is the subset of T with first coordinate x; for $y^D \leqq y \leqq y^U$, the yth *row of* T is defined analogously. By convexity, each row and each column consists of a set of consecutive squares. For each $x^L \leqq x \leqq x^R$, we define $U(x) = \max\{y'|(x, y') \in T\}$ and $D(x) = \min\{y'|(x, y') \in T\}$, thus the xth column of T equals $[x]\times[D(x), U(x)]$. Similarly, for $y^D \leqq y \leqq y^U$ we have $R(y) = \max\{x'|(x', y) \in T\}$ and $L(y) = \min\{x'|(x', y) \in T\}$, so the yth row of T equals $[L(y), R(y)]\times[y]$.

A sequence $\{a_i|i \in \mathbb{Z}, m \leqq i \leqq n\}$ is *unimodal* if there exists $m \leqq k \leqq n$ such that $a_i \leqq a_{i+1}$ for $i < k$ and $a_i \geqq a_{i+1}$ for $i \geqq k$. The following proposition provides an equivalent condition for convexity (the proof is omitted).

PROPOSITION 2.1. *T is convex if and only if it is simply connected (in the topological sense) and the sequences $(U(x)|x^L \leqq x \leqq x^R)$, $(-D(x)|x^L \leqq x \leqq x^R)$, $(R(y)|y^D \leqq y \leqq y^U)$ and $(-L(y)|y^D \leqq y \leqq y^U)$ are each unimodal.*

The set of squares of T lying in row y^U (respectively, y^D) is called the *upper* (respectively, *lower*) *boundary support* of T and the set lying in column x^R (respectively, x^L) is called the *right* (respectively, *left*) *boundary support* of T.

If $s \neq t \in T$ and some rectangle of T contains them, we say *s is related to t* and write $s \sim t$. A square is not related to itself. If (x_1, y_1) and (x_2, y_2) are related then the opposite corners (x_1, y_2) and (x_2, y_1) must be in T, and convexity implies that this is sufficient. In fact we can strengthen this to:

PROPOSITION 2.2. *Suppose $s_1 = (x_1, y_1)$ and $s_2 = (x_2, y_2)$ belong to the convex connected region T with s_1 above and to the left of s_2. Then $s_1 \sim s_2$ if and only if T contains a square s_3 lying below and left of both s_1 and s_2 and a square s_4 lying above and to the right of both s_1 and s_2. (By symmetry, the result holds if "above" and "below" are interchanged throughout.)*

Proof. "Only if" is obvious since (x_1, y_2) and (x_2, y_1) satisfy the conditions of s_3 and s_4. Conversely, given s_3 and s_4 as stipulated there are paths from s_3 to s_1 and s_2 and from s_4 to s_1 and s_2 by connectivity. Together with convexity this implies (x_1, y_2) and (x_2, y_1) are in T. □

If $s = (x, y)$ is a square we define RECT(s) to be the set $[L(y), R(y)]\times[D(x), U(x)]$ (see Fig. 2.1). RECT(s) is generally not in T. The following is an immediate consequence of the previous proposition.

PROPOSITION 2.3. *If $s, t \in T$ then $s \sim t$ if and only if $t \in$* RECT(s).

If T is a region and $F \subseteq T$, the *neighborhood in F* of a square $s \in T$, written $N_F(s)$, is the set $\{t|s \sim t, t \in F\}$. By Proposition 2.3, $N_F(s) = F \cap$ RECT$(s)\backslash s$.

The next proposition justifies the assertion made in the introduction that maximal cliques in G^T correspond to maximal rectangles in T.

PROPOSITION 2.4. *Let $s_1 = (x_1, y_1), s_2 = (x_2, y_2), \cdots, s_n = (x_n, y_n)$ be a set of squares in T such that $s_i \sim s_j$ if $i \neq j$. Then there exists a rectangle R of T with $\{s_1, \cdots, s_n\} \subseteq R$.*

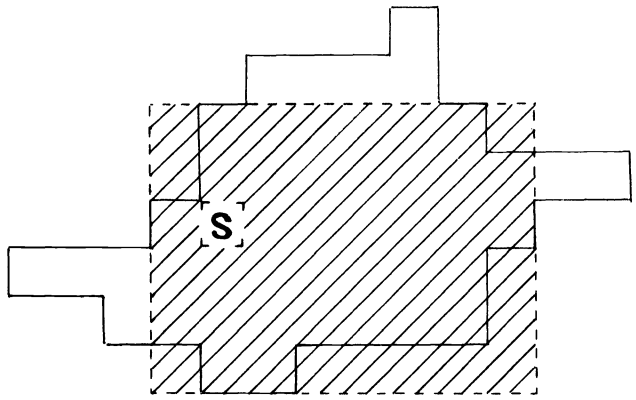


FIG. 2.1. RECT (*s*) *is shaded.*

Proof. Let $x_{\min}$, $x_{\max}$, $y_{\min}$, and $y_{\max}$ be the minimum and maximum x and y coordinates of any square s_i. Since $s_i \sim s_j$, we have $[\min(x_i, x_j), \max(x_i, x_j)] \times [\min(y_i, y_j), \max(y_i, y_j)]$ is contained in T for all $0 \leqq i, j \leqq n$. Taking the union of all these rectangles yields the rectangle $[x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}]$ (as is easily verified), which is the desired rectangle. □

3. Proof of Theorem 1.1. The proof of Theorem 1.1 is by induction. We order all boards (T, F) (where T is convex and F consists of border squares), by $(T', F') \leqq (T, F)$ if $|T'| \leqq |T|$ and $|F'| \leqq |F|$. The basis step of the induction is trivial. Let (T, F) be a board and assume the theorem holds for all smaller boards.

The induction step relies on a set of nine *reductions*. Each reduction has three parts:

(1) The *conditions* which (T, F) must satisfy for the reduction to apply.

(2) The *construction* in which a smaller board (T', F') is constructed.

(3) By induction (T', F') has an antirectangle A' and an equal-sized cover $\mathscr{C}'$. The *correspondence* describes how to use A' and $\mathscr{C}'$ to obtain an antirectangle A and an equal sized cover $\mathscr{C}$ of (T, F).

If the board (T, F) satisfies the conditions of a reduction (that is, is *reducible*), then by induction and the correspondence the theorem holds for (T, F). The proof of the theorem consists of showing that every nontrivial board is reducible.

Once a reduction has been presented we can assume, for purposes of subsequent analysis, that (T, F) does not meet the conditions of that reduction.

By symmetry, each reduction we present corresponds to several reductions. For example, if the condition for a reduction is that the lower left corner has a certain property, then if an analogous condition holds for an upper right corner then the board is reducible.

Reduction 1.

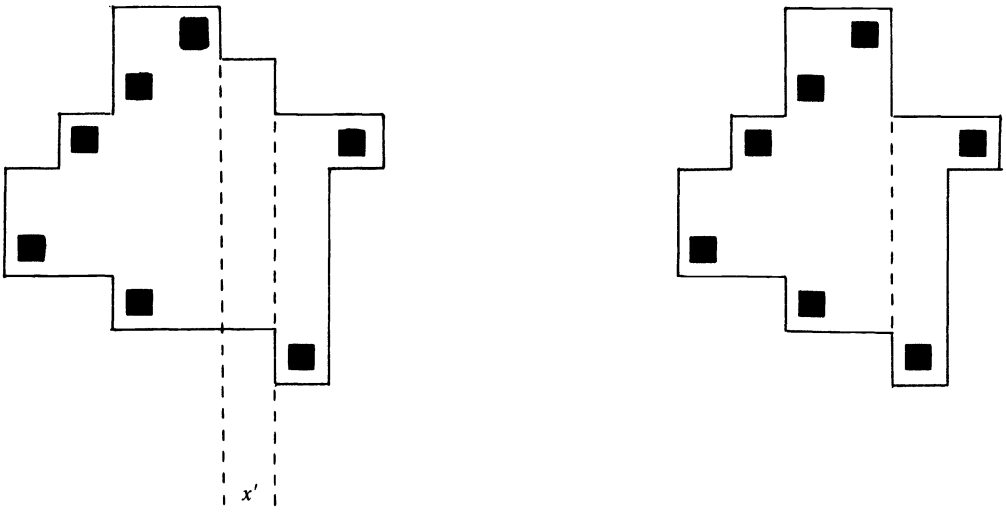
Condition. There exists a column of T containing no square of F. Let x' be the coordinate of the column (Fig. 3.1(i)).

Construction. T' and F' are obtained by deleting column x' and shifting all squares lying to the right of that column to the left by one unit (Fig. 3.1(ii)).

Correspondence. A is obtained from A' by reversing the above shift. $\mathscr{C}$ is obtained by replacing each rectangle $[x_1, x_2] \times [y_1, y_2]$ in $\mathscr{C}'$ by:

$$\begin{aligned} [x_1, x_2] \times [y_1, y_2] \quad & \text{if } x_2 < x', \\ [x_1, x_2 + 1] \times [y_1, y_2] \quad & \text{if } x_1 < x' \leqq x_2, \\ [x_1 + 1, x_2 + 1] \times [y_1, y_2] \quad & \text{if } x' \leqq x_1. \end{aligned}$$

These rectangles obviously cover F. In the first and third case, it is clear they lie in T; in the second case convexity implies that the situation depicted in Fig. 3.2 cannot happen, so the rectangle is in T.



(i) Condition. No square of F in column x'.

(ii) Construction. Delete column x'.

FIG 3.1. *Reduction* 1.

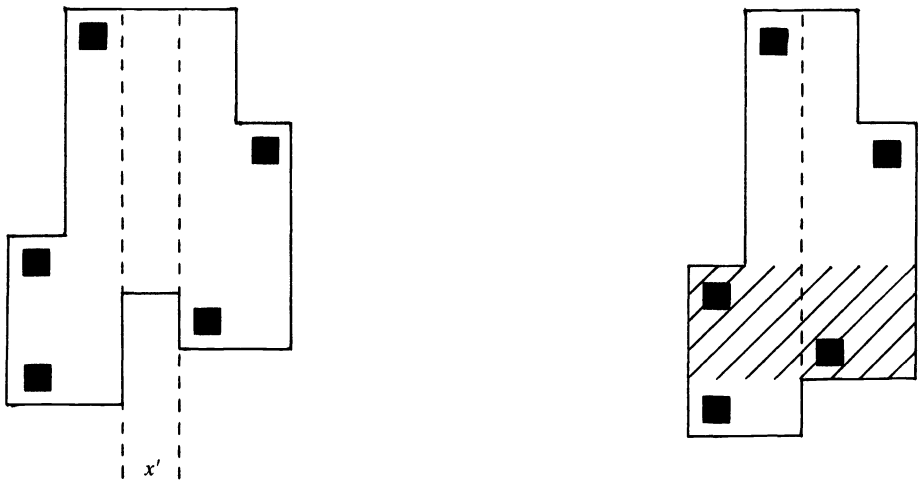


FIG 3.2. *The shaded rectangle in the reduced board (at right) does not correspond to a rectangle in the original board; however, convexity precludes this situation.*

Reduction 2.

Condition. Two adjacent columns have the same upper and lower boundaries, i.e., there exists x' such that $U(x') = U(x'+1)$ and $D(x') = D(x'+1)$ (Fig. 3.3(i)).

Construction. Merge columns x' and $x'+1$. Formally, let

$$T' = \{(x, y) | (x, y) \in T, x \leqq x' \text{ or } (x+1, y) \in T, x \geqq x'\},$$

and

$$F' = \{(x, y) | (x, y) \in F, x \leqq x' \text{ or } (x+1, y) \in F, x \geqq x'\},$$

(see Fig. 3.3(ii)).

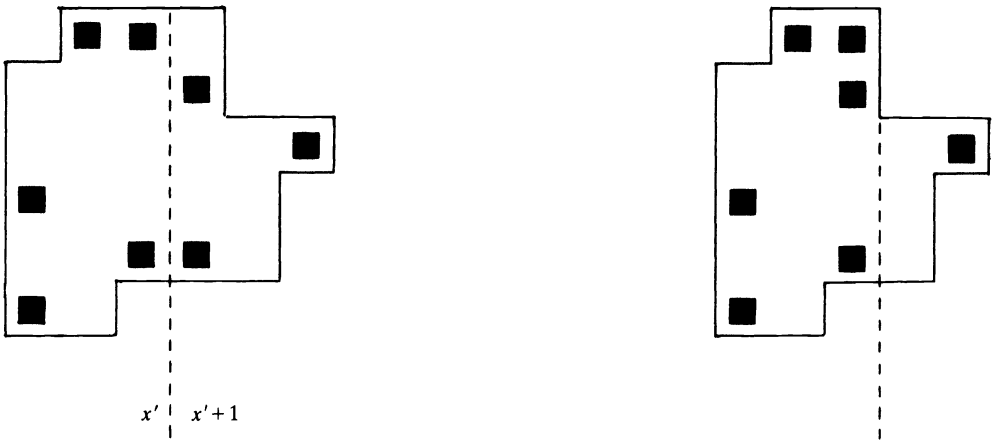
Correspondence. A and $\mathscr{C}$ are obtained from A' and $\mathscr{C}'$ by reversing the above construction.

Reduction 3.

Condition. There is only one square in the uppermost row (y^U) of T. Let $s = (x', y^U)$ be the square. By RED1, we can assume $s \in F$ (Fig. 3.4).

Construction. $T' = T$; $F' = \{(x, y) \in F | x \neq x'\}$ (Fig. 3.4(ii)).

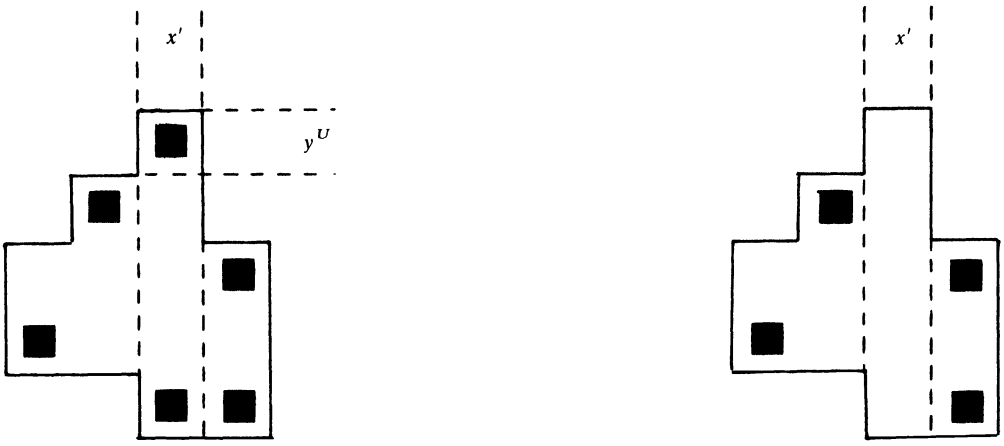
Correspondence. $A = A' \cup s$ is an antirectangle of (T, F) with $|A| = |A'| + 1$. $\mathscr{C}$ is obtained by adding rectangle $[x'] \times [D(x'), y^U]$ to $\mathscr{C}'$.



(i) Condition. Columns x' and $x'+1$ have the same upper and lower boundaries.

(ii) Construction. Merge columns x' and $x'+1$.

FIG. 3.3. *Reduction* 2.



(i) Condition. Only one square of T in row y^U.

(ii) Construction. Delete all squares in F from column x'.

FIG 3.4. *Reduction* 3.

Reduction 4.

Condition. There exists $s, t \in F$ with $s \sim t$, and everything else in F which is related to t is also related to s, i.e., $N_F(t) \subseteq N_F(s) \cup s$.

Construction. $T' = T$; $F' = F \backslash s$.

Correspondence. $A = A'$. Let R' be the rectangle in $\mathscr{C}'$ covering t. By the reduction condition, $R' \cap F \subseteq N_F(s) \cup s$ so by Proposition 2.4 there is a rectangle R covering $(R' \cap F) \cup s$. $\mathscr{C}$ is obtained by replacing R' by R in $\mathscr{C}$.

Reduction 5.

Condition. Two squares in F lie on the same boundary segment. Suppose $s = (x, y)$ and $t = (x, y')$ lie on a left boundary segment (in column x), and assume $R(y) \supseteqq R(y')$. Then RECT $(t) \subseteq$ RECT (s) and so by Proposition 2.3 we can apply RED4 (Fig. 3.5).

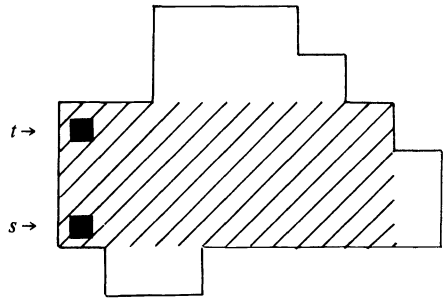


FIG. 3.5. *Condition for reduction* 5. *t and s lie on the same boundary segment and* RECT $(t) \subseteq$ RECT (s). RECT (t) *is shaded.*

Reduction 6.

Condition. There exists a corner $(x, y) \notin F$ such that no pair of squares $(x', y), (x, y') \in F$ are related.

Construction. $T' = T \backslash (x, y)$ and $F' = F$.

Correspondence. $A = A'$ is an antirectangle in T since, under the given conditions, deleting (x, y) does not alter any relations among squares in F. $\mathscr{C} = \mathscr{C}'$ is a cover of (T, F).

Reduction 7.

Condition. There exist squares $s, s' \in F$ such that

(i) $s \sim s'$;

(ii) $N_F(s) \cap N_F(s') = \varnothing$;

(iii) if $t, t' \in F$ and $s \sim t$ and $s' \sim t'$ then $t \sim t'$ (Fig. 3.6).

Construction. $T' = T$; $F' = F \backslash s, s'$.

Correspondence. By condition (iii), A' cannot contain both a square in $N_F(s)$ and a square in $N_F(s')$, hence either $A' \cup s$ or $A' \cup s'$ is an antirectangle A in (T, F) of size one larger than A'. By adding a rectangle that covers s and s' to $\mathscr{C}'$, we obtain $\mathscr{C}$ with $|\mathscr{C}| = |A|$.

Reduction 8.

Condition. There exist squares $t_1, t_2 \in F$ and $t_3 \in T \backslash F$ such that:

(i) $t_1 \sim t_2$;

(ii) $t_1 \sim t_3$, $t_2 \sim t_3$;

(iii) $N_F(t_3) \backslash t_1, t_2 \subseteq N_F(t_1) \cap N_F(t_2)$;

(iv) if $u_1 \in N_F(t_1) \backslash N_F(t_3)$ and $u_2 \in N_F(t_2) \backslash N_F(t_3)$ then $u_1 \neq u_2$ and $u_1 \sim u_2$ (Fig. 3.7).

Construction. $T' = T$; $F' = F \cup t_3 \backslash t_1, t_2$.

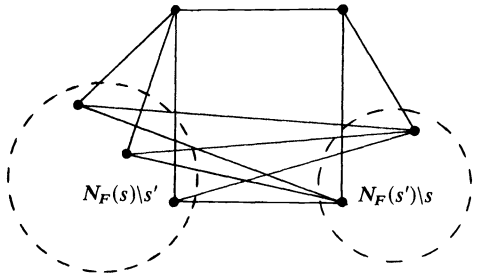


FIG. 3.6. *Condition for reduction* 7, *which depends only on the structure of the graph* G_F^T.

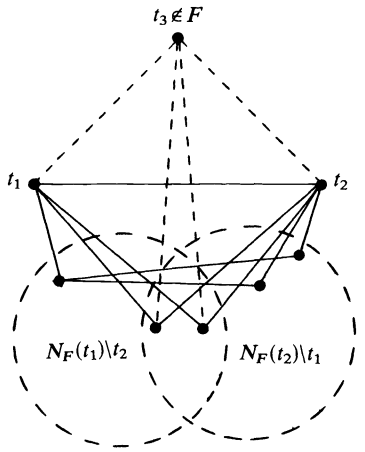


FIG. 3.7. *Condition for reduction* 8, *which depends only on the structure of* G^T.

Correspondence. If $t_3 \notin A'$, then $A = A'$. If $t_3 \in A'$ then either $A' \cup t_1 \backslash t_3$ or $A' \cup t_2 \backslash t_3$ is an antirectangle since, by (iv), A' cannot contain both an element belonging to $N_F(t_1) \backslash N_F(t_3)$ and an element belonging to $N_F(t_2) \backslash N_F(t_3)$.

Let R' be the rectangle in $\mathscr{C}'$ covering t_3. By (iii), every square in $R' \cap F$ is related to t_1 and t_2 and thus, by Proposition 2.4, there is a rectangle R of T containing $(R' \cap F) \cup \{t_1, t_2\}$. Let $\mathscr{C} = \mathscr{C}' \cup R \backslash R'$.

Reduction 9.

Condition. There exist squares $s_1 = (x_1, y_1)$ and $s_2 = (x_2, y_2)$ in F, $s_1 \sim s_2$, such that neither lie above the left or right boundary supports and neither $R(x_1)$ nor $R(x_2)$ lie to the right of the right corner of the upper boundary support. Formally,

(i) $y_1, y_2 \leqq U(x^L), U(x^R)$,

(ii) $R(y_1), R(y_2) \leqq R(y^U)$

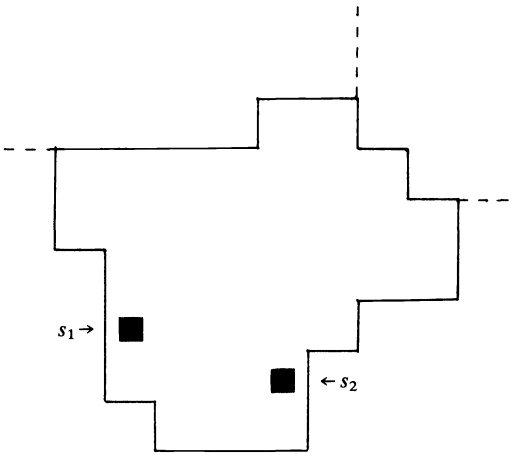
(Fig. 3.8(i)). It is clear that if s_1 and s_2 satisfy these conditions, then any pair of related squares with smaller y components satisfy them, so we assume s_1 and s_2 have minimal y components, i.e., for any other pair of related squares (x_1', y_1') and (x_2', y_2'), if $y_1' \leqq y_1$ and $y_2' \leqq y_2$, then $y_1' = y_1$ and $y_2' = y_2$. We label the squares so that $y_2 \leqq y_1$ and if $y_1 = y_2$ then $x_1 > x_2$.

There are two cases.

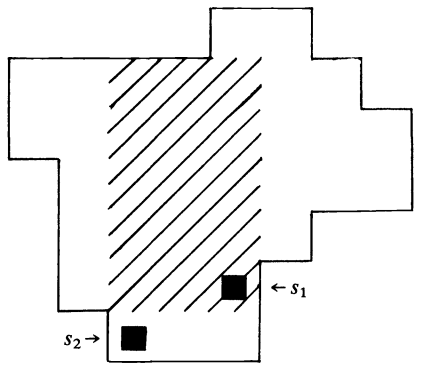
Case 1. $x_2 \leqq x_1$. We show that if $t \in F \backslash s_1$ and $t \sim s_2$ then $t \sim s_1$ and thus RED4 applies. By Proposition 2.3 we have $t \in [L(y_2), R(y_2)] \times [D(x_2), U(x_1)]$. By condition (i) and convexity, $[L(y_2), R(y_2)] \subseteq [L(y_1), R(y_1)]$. By the minimality of s_1 and s_2, t cannot have a y-component smaller than y_1 or else $\{s_2, t\}$ would be a lower pair of

related squares. By (ii) and convexity, $U(x_2) \leqq U(x_1)$ so we conclude $t \in [L(y_1), R(y_1)] \times [D(x_1), U(x_1)]$ and thus $t \sim s_1$ (Fig. 3.8(ii)).

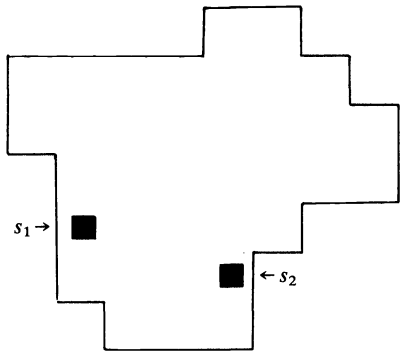
Case 2. $x_2 > x_1$. By hypothesis, $y_1 > y_2$. Since $s_1 \sim s_2$, s_1 lies on a left boundary segment and $D(x_1) \leqq y_2$ (Fig. 3.8(iii)). If $D(x_1) < y_2$ then RED6 can be applied to delete corner $(x_1, D(x_1))$, therefore (x_1, y_2) is a corner. Now we show that the conditions of RED8 apply with $t_3 = (x_1, y_2)$, $t_2 = s_2$ and $t_1 = s_1$. The three squares are related, so conditions (i) and (ii) hold. Examine RECT (t_1), RECT (t_2) and RECT (t_3) (Fig. 3.8(iv)). We have RECT (t_3) = RECT $(t_1) \cap$ RECT (t_2) (by convexity and the conditions of RED9) so condition RED8(iii) holds. To verify RED8(iv) first note that by the minimality of s_1 and s_2, nothing related to s_2 lies strictly below s_1 (in rectangle A_1 of Fig. 3.8(iv)). Thus $N_F(s_1) \backslash N_F(x_1, y_2) \subseteq A_2$ (Fig. 3.8(iv)) and $N_F(s_2) \backslash N_F(x_1, y_2) \subseteq A_3$, so it suffices to show that each square in A_2 is related to each square in A_3. Obviously (x_1, y_2) lies below and to the left of all such squares and by condition RED9(ii), the square $(R(y^U), y^U)$ is above and to the right of all such squares. By Proposition 2.2, therefore, everything in $N_F(s_1) \backslash N_F(x_1, y_2)$ is related to everything in $N_F(s_2) \backslash N_F(x_1, y_2)$ as required for condition (iv) of RED8.



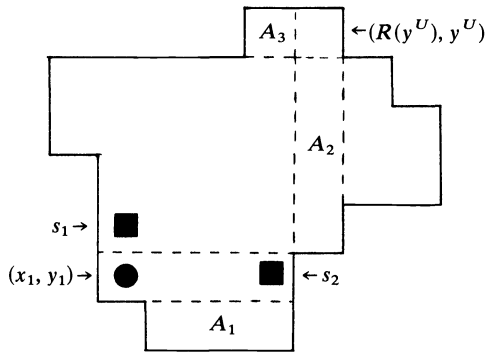
(i) Condition.



(ii) If $x_2 \leqq x_1$, then RECT $(s_2) \cap F$ is contained in the shaded rectangle which is contained in RECT $(s_1) \cap F$.



(iii) If $x_2 > x_1$, then s_1 lies on a left boundary segment.



(iv) The conditions of RED8 hold between s_1, s_2 and (x_1, y_2).

FIG. 3.8. *Reduction* 9.

We are now ready to show that every board contains a reducible configuration which will complete the proof of Theorem 1.1. Suppose (T, F) is not reducible; we use irreducibility to get a precise description of the board and eventually get a contradiction. We proceed by a series of numbered observations.

Fact 1. Consider the squares lying on the upper and lower boundary supports, $B^U = [L(y^U), R(y^U)] \times [y^U]$ and $B^D = [L(y^D), R(y^D)] \times [y^D]$. By RED3, $L(y^U) \neq R(y^U)$ and $L(y^D) \neq R(y^D)$ and by RED2, $[L(y^U), R(y^U)]$ and $[L(y^D), R(y^D)]$ overlap in at most one point, so without loss of generality assume $R(y^D) \leqq L(y^U)$. By convexity, $D(x)$ is nondecreasing for $x \geqq R(y^D)$ and so by RED2, $D(x)$ is strictly increasing for $L(y^U) \leqq x \leqq R(y^U)$ and thus the lower boundary opposite B^U forms a "rising staircase" (Fig. 3.9).

Fact 2. By RED1 and RED5 there is a unique square $s^U = (x^U, y^U)$ in F which lies on B^U. If s^U is related to some square $s \in F$ lying to the left of it, then $N_F(s^U) \subseteq N_F(s) \cup s$, so RED4 applies (Fig. 3.10). Now we claim that s^U must lie in the left corner of B^U, for if not, that corner satisfies the conditions of RED6. Hence $x^U = L(y^U)$.

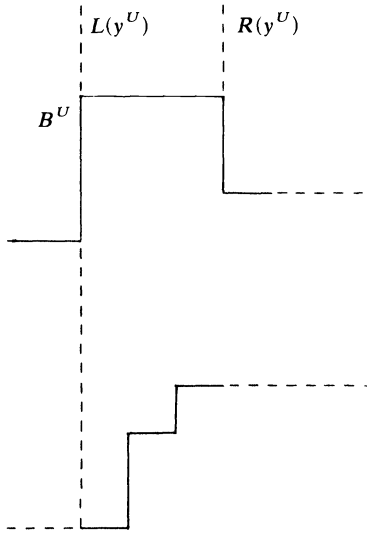


FIG. 3.9. *Fact* 1. $D(x)$ *is strictly increasing for* $L(y^U) \leqq x \leqq R(y^U)$.

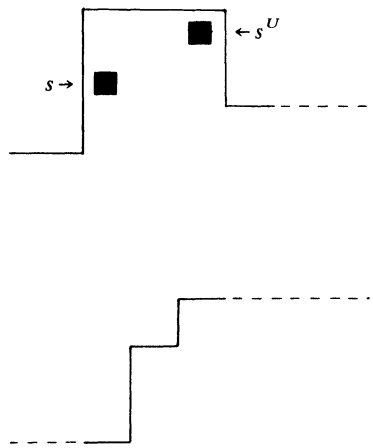


FIG. 3.10. *If* s^U *is related to some square* $s \in F$ *lying to the left of it then* RED4 *applies.*

Fact 3. The square $(x^U, D(x^U))$ is a lower right corner. The right boundary segment incident to it extends up to $D(x^U+1)-1$. There must be a square $s_1=(x^U, y_1)$ in F which lies on that boundary segment or else the corner $(x^U, D(x^U))$ satisfies the conditions of RED6; by RED5 the square is unique. Now by RED1, column $R(y^U)$ has a square in F; let $s_2=(x_2, y_2)$ be the square of highest y component in that column. Notice every square in F that is related to s^U lies on the "rising staircase" between s_1 and s_2 (Fig. 3.11). By Proposition 2.3, any square related to both s_1 and s_2 is related to every square between them and thus to every square in $N_F(s^U)$.

Fact 4. The squares (besides s^U) related to s_1 lie to the left of column x^U and above or even with $D(x^U)$ (Fig. 3.12).

Fact 5. By Fact 3, if every square in F which is related to s_1 is also related to s_2, then they are related to every square in $N_F(s^U)$, in which case, s_1 and s^U satisfy the conditions for RED7. Thus there is a square $s_3=(x_3, y_3)$ in F which is related to s_1 but not s_2. If s_3 was above s_2 they would be related, so s_3 must lie strictly below s_2 (Fig. 3.13).

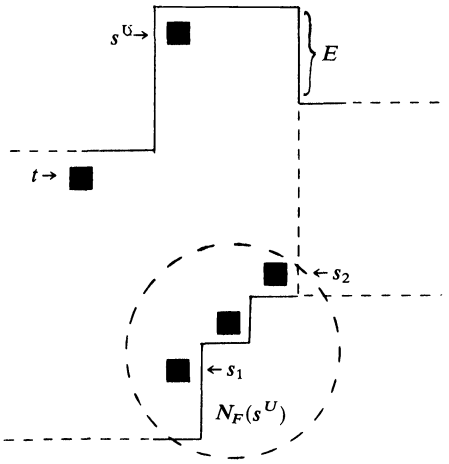


FIG. 3.11. *Fact* 3. *Any square t related to both* s_1 *and* s_2 *is related to every square in* $N_F(s^U)$. (*Note that* s_2 *might be along the edge E rather than where it is shown here*).

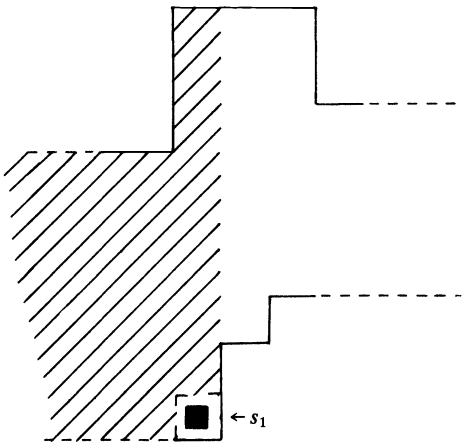


FIG. 3.12. *Fact* 4. $N_F(s_1)$ *consists of squares above and to the left of* s_1.

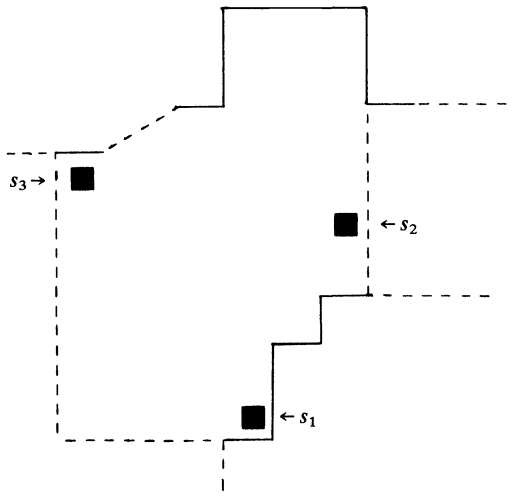


FIG. 3.13. *Fact* 5. *If* $s_3 \sim s_1$ *and* s_3 *does not lie strictly below* s_2 *then* $s_2 \sim s_3$.

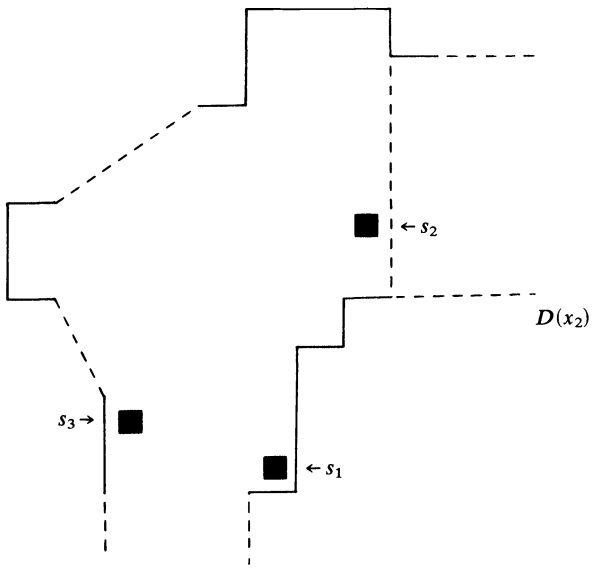


FIG. 3.14. *Fact* 6. *If* $y_3 < D(x_2)$ *and the upper edge of the left boundary support lies above* s_1 *and* s_2 *then* RED9 *applies.*

Fact 6. If $y_3 < D(x_2)$ then the upper edge of the left boundary support cannot lie above both s_1 and s_3 (Fig. 3.14) or else s_1 and s_3 would satisfy the conditions of RED9.

Fact 7. s_2 is not related to any square lying strictly to the right of it. Suppose, to the contrary, that $s_4 = (x_4, y_4)$ were such a square. Then s_2 cannot lie along a right boundary segment and, therefore, must be the square $(x_2, D(x_2))$. By Fact 5, s_3 lies below s_2 and thus by Fact 6, the upper edge of the left boundary support does also. Now by reflecting the board around the line $y = -x$, RED9 applies to s_2 and s_4 (Fig. 3.15).

Fact 8. $L(y_1) < L(y_2)$ (Fig. 3.16). If instead $L(y_1) \geqq L(y_2)$ then since s_3 is related to s_1 but not s_2 we must have $y_3 < D(x_2)$ and by convexity, the upper edge of the left boundary support lies above both s_1 and s_3 contrary to Fact 6 (Fig. 3.17).

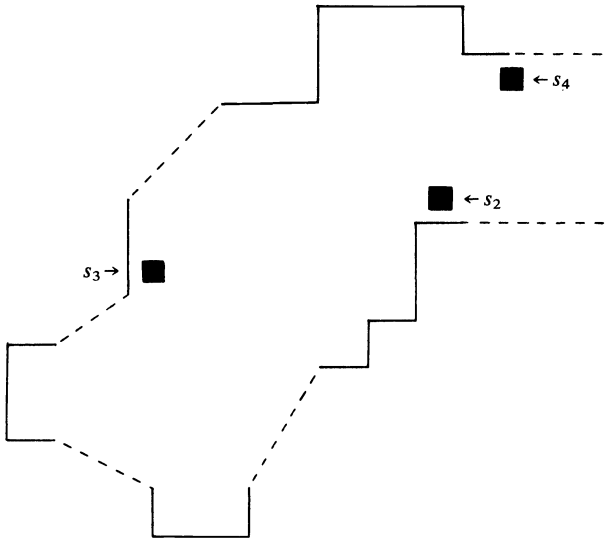


FIG. 3.15. *Fact* 7. RED9 *applies to* s_2 *and* s_4 *when board is rotated around the line* $y = -x$.

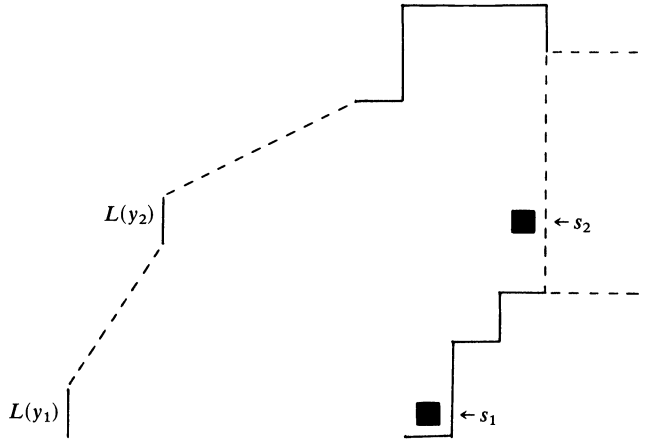


FIG. 3.16. *Fact* 8. $L(y_1) < L(y_2)$.

Fact 9. Now RED8 applies to $t_1 = s^U$, $t_2 = s_2$ and $t_3 = (R(y^U), y^U)$ (Fig. 3.18). Conditions (i), (ii) and (iii) of RED8 are obvious. By Fact 7, every square in $N_F(s_2) \backslash N_F(R(y^U), y^U)$ lies to the left of s^U. By Fact 8, every such square is related to s_1 and thus, by Fact 3, to every square in $N(s^U)$. Hence condition RED8(iv) is met.

This completes the proof of Theorem 1.1. □

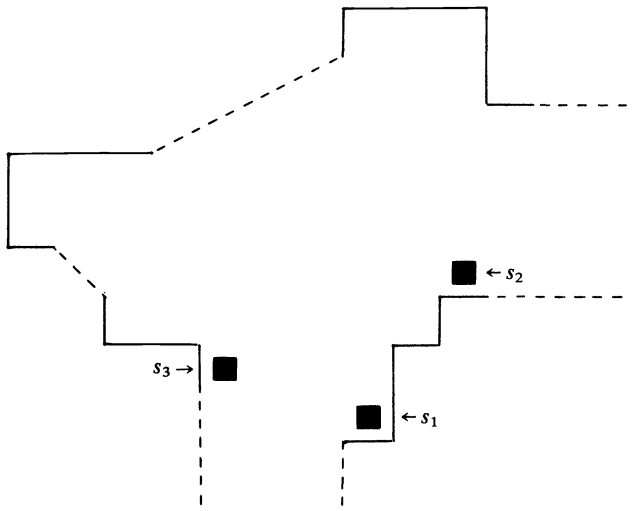


FIG. 3.17. *Fact* 8. *If* $L(y_1) \geqq L(y_2)$ *then the upper edge of the left boundary support lies above* s_1 *and* s_3, *contrary to Fact* 6.

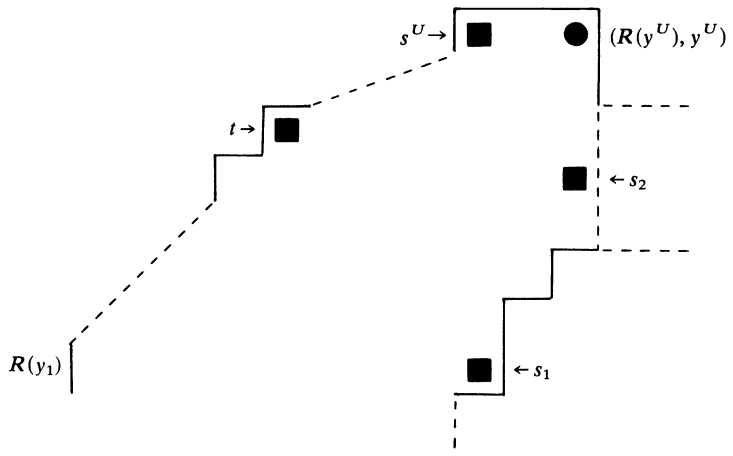


FIG. 3.18. *Fact* 9. RED8 *applies to* $t_1 = s^U$, $t_2 = s_2$, $t_3 = (R(y^U), y^U)$. *If* $t \sim s_2$ *and* $t \not\sim s^U$ *then* $t \sim s_1$.

REFERENCES

[1] C. BERGE, *Graphs and Hypergraphs*, North-Holland, Amsterdam, 1976.
[2] S. CHAIKEN, D. J. KLEITMAN, M. SAKS AND J. SHEARER, *Covering regions by rectangles*, this Journal, 2 (1981), pp. 394–410.
[3] M. GOLUMBIC, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.
[4] W. MASEK, *Some* NP-*complete set covering problems*, MIT Laboratory of Computer Science Technical Memo, Massachusetts Institute of Technology, Cambridge, 1980.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 3, September 1982

0196-5212/82/0303-0009 $01.00/0

THE STRUCTURE OF HOMOMETRIC SETS*

JOSEPH ROSENBLATT† AND PAUL D. SEYMOUR‡

Abstract. One of the fundamental problems of phase retrieval in spectroscopic analysis is of a combinatorial nature and can be solved using purely algebraic techniques. Given two sets A and B in some Euclidean space R^n, A and B are *homometric* if the sets of vector differences $\{x-y: x, y \in A\}$ and $\{x-y: x, y \in B\}$ are identical counting multiplicities. More generally, given two finite sums $A=\sum a_x\delta_x$ and $B=\sum b_x\delta_x$, where a_x, b_x are integers and δ_x denotes the Dirac mass at $x \in R^n$, A and B are *homometric* if they have the same Patterson functions, i.e., for all $z \in R^n$, $\sum\{a_xa_y: x-y=z\}=\sum\{b_xb_y: x-y=z\}$. Using a variation on factorization of polynomials with integer coefficients, one can prove that A and B are homometric if and only if there exists two finite sums $C=\sum c_x\delta_x$ and $D=\sum d_x\delta_x$ such that A is the convolution $C*D$ and B is the convolution $C*D^*$, where $D^*(x)=D(-x)$ for all $x \in R^n$. Moreover, the algebraic method above allows one to derive both practically and theoretically, from the Patterson function $A*A^*$, all sums B with A and B homometric.

1. Introduction. The following combinatorial problem arises naturally in spectroscopic analysis of matter. If $A \subseteq R^n$ is a multiset (a finite set with repetitions allowed), let ΔA denote the multiset of all vector differences $x-y$ with $x, y \in A$. The problem is to reconstruct A as far as possible from a knowledge of ΔA alone. In § 3, we will explain the connection of this problem with the problem of phase retrieval in spectroscopic analysis, i.e., retrieval of the distribution determined by A from a knowledge of the intensity of the Fourier transform of this distribution.

It is not possible to completely reconstruct A from a knowledge of ΔA. For example, for any $v \in R^n$, with $A+\{v\}$ being the multiset $\{a+v: a \in A\}$, we have $\Delta A=\Delta(A+\{v\})$. Also, if $-A=\{-a: a \in A\}$, then $\Delta(A)=\Delta(-A)$, too.

There are less trivial examples of this nonuniqueness. Let us say that two multisets A and B are *homometric* if $\Delta A=\Delta B$. The two constructions above are special cases of the following. Let $U, V \subset R^n$ be two multisets. Then the multisets

$$U+V=\{u+v: u \in U, v \in V\}$$

and

$$U-V=\{u-v: u \in U, v \in V\}$$

are homometric, as one can see very easily. Thus, for example, the sets $\{0, 1, 3, 8, 9, 11, 12, 13, 15\}$ and $\{0, 1, 3, 4, 5, 7, 12, 13, 15\}$ in R are homometric and arise from the above construction by taking $U=\{6, 7, 9\}$ and $V=\{-6, 2, 6\}$.

It is natural to ask if every homometric pair is an instance of this construction. The answer is negative; for example, one can check that $\{0, 1, 2, 5, 7, 9, 12\}$ and $\{0, 1, 5, 7, 8, 10, 12\}$ are homometric but do not arise from the above construction. Indeed, if they did arise this way, then the multiplicities $|U|$ and $|V|$ would satisfy $|U||V|=|U+V|=7$ and either $|U|$ or $|V|$ equals 1, an impossibility. Nevertheless, this conjecture is true if we permit U and V to be "multisets with possibly negative integer multiplicities". Moreover, this representation is accomplished by a purely algebraic technique.

2. The main result. Let K be one of the rings of integers, Z, real numbers, R, or complex numbers, C, under addition. The examples of the introduction only require

* Received by the editors June 30, 1981.

† Department of Mathematics, Ohio State University, Columbus, Ohio 43210. The work of this author was partially supported by the National Science Foundation under grant MCS 8002881.

‡ Department of Mathematics, Ohio State University, Columbus, Ohio 43210.

$K = Z$. Let n be a whole number, the dimension of the Euclidean space R^n from which the multiset is chosen. Let $x_1, \cdots, x_n$ be some n commuting variables and let $K[R^n]$ be the ring under multiplication of all sums

$$A(x_1, \cdots, x_n) = \sum \{a_{v_1,\cdots,v_n} x_1^{v_1} \cdots x_n^{v_n} : (v_1, \cdots, v_n) \in R^n\}$$

where $a_{v_1 \cdots v_n} \in K$ and $a_{v_1 \cdots v_n} \neq 0$ for only finitely many choices of $(v_1, \cdots, v_n) \in R^n$. It is convenient to abbreviate $(v_1, \cdots, v_n)$ by v and $(x_1, \cdots, x_n)$ by x, and $x_1^{v_1} \cdots x_n^{v_n}$ by x^v. Then a typical element $A(x) \in K[R^n]$ may be written more compactly as

$$A(x) = \sum \{a_v x^v : v \in R^n\}.$$

The ring operation in $K[R^n]$ is the usual multiplication with $x^v x^w = x^{v+w}$ for all $v, w \in R^n$. The reader familiar with group rings will see that $K[R^n]$ is just the group ring of R^n with coefficients in K. We shall need the following theorem, which is well known.

THEOREM 2.1. *With $K = Z$, R, or C, the ring $K[R^n]$ is locally a unique factorization domain. The units in $K[R^n]$ are the elements ux^v where $v \in R^n$ and u is a unit of K.*

Remark. Here we mean that if $A_1, \cdots, A_k \in K[R^n]$, there exists $Z^m \in R^n$ such that $A_1, \cdots, A_k \in K[Z^m]$ and $K[Z^m]$ is a unique factorization domain. All factoring arguments below will be local to some suitable $K[Z^m]$ in this sense.

Given $A \in K[R^n]$, $A(x) = \sum a_v x^v$, we let $A(x^{-1}) = \sum \bar{a}_v x^{-v}$, where $\bar{a}_v$ is the complex conjugate of a_v. We say that two elements $A_1(x)$ and $A_2(x)$ in $K[R^n]$ are *homometric* if $A_1(x)A_1(x^{-1}) = A_2(x)A_2(x^{-1})$. The motivation for this is as follows. If $A \subseteq R^n$ is a multiset, we associate with it the element $A(x) \in Z[R^n]$ defined by $A(x) = \sum \{x^v : v \in A\}$. Then $A(x)A(x^{-1}) = \sum \{x^{v-w} : v, w \in R^n\} = \sum \{x^v : v \in \Delta A\}$. Hence, A_1 and A_2 are homometric as multisets if and only if $A_1(x)$ and $A_2(x)$ are homometric as ring elements.

Our main result is the following.

THEOREM 2.2. *Two elements $A_1(x)$, $A_2(x)$ in $K[R^n]$ are homometric if and only if there exist $P(x)$, $Q(x)$ in $K[R^n]$ and a member $c \in K$ of absolute value 1 such that $A_1(x) = P(x)Q(x)$ and $A_2(x) = cP(x)Q(x^{-1})$.*

Proof. Certainly, if $P(x)$, $Q(x) \in K[R^n]$, then $P(x)Q(x)$ and $cP(x)Q(x^{-1})$ are homometric, as long as $c\bar{c} = 1$, because $P(x)Q(x)(P(x^{-1})Q(x^{-1})) = P(x)Q(x^{-1})P(x^{-1})Q((x^{-1})^{-1})$. Conversely, suppose $A_1(x), A_2(x) \in K[R^n]$ are homometric. Because $K[R^n]$ is locally a unique factorization domain, we can write

$$A_1(x) = P_0(x)B_1(x), \qquad A_2(x) = P_0(x)B_2(x),$$

where $P_0(x), B_1(x), B_2(x) \in K[R^n]$ and $B_1(x), B_2(x)$ are relatively prime. Now write

$$B_1(x) = Q_0(x)C_1(x), \qquad B_2(x^{-1}) = Q_0(x)C_2(x),$$

where $Q_0(x), C_1(x), C_2(x) \in K[R^n]$ and $C_1(x), C_2(x)$ are relatively prime. Then because $A_1(x)A_1(x^{-1}) = A_2(x)A_2(x^{-1})$, we have

$$P_0(x)Q_0(x)C_1(x)P_0(x^{-1})Q_0(x^{-1})C_1(x^{-1}) = P_0(x)Q_0(x)C_2(x)P_0(x^{-1})Q_0(x^{-1})C_2(x^{-1}).$$

So

$$C_1(x)C_1(x^{-1}) = C_2(x)C_2(x^{-1}). \tag{1}$$

Moreover, both $C_1(x)$ and $C_1(x^{-1})$ are relatively prime to both $C_2(x)$ and $C_2(x^{-1})$.

Again, because $K[R^n]$ is locally a unique factorization domain, any prime $D(x)$ which divides $C_1(x)$ must divide $C_2(x)C_2(x^{-1})$ and hence divides $C_2(x)$ or $C_2(x^{-1})$. This is impossible and so $C_1(x)$ is a unit. Similarly, $C_2(x)$ is a unit. By Theorem 2.1, there are units $u_1, u_2 \in K$ and vectors $v_1, v_2 \in R^n$ with $C_1(x) = u_1 x^{v_1}$, $C_2(x) = u_2 x^{v_2}$. By

(1) $u_1\bar{u}_1 = u_2\bar{u}_2$ and hence there exists $c \in K$, $c\bar{c} = 1$, such that $u_1 = c\bar{u}_2$. Put $j_1 = \frac{1}{2}(v_1 - v_2)$ and $j_2 = \frac{1}{2}(v_1 + v_2)$. Define $P(x) = \bar{u}_2 x^{j_1} P_0(x)$ and $Q(x) = cx^{j_2} Q_0(x)$. Then we have

$$P(x)Q(x) = c\bar{u}_2 x^{j_1+j_2} P_0(x)Q_0(x) = u_1 x^{v_1} P_0(x)Q_0(x)$$
$$= C_1(x)P_0(x)Q_0(x) = A_1(x),$$

and

$$cP(x)Q(x^{-1}) = c\bar{c}\bar{u}_2 x^{j_1-j_2} P_0(x)Q_0(x^{-1}) = \bar{u}_2 x^{-v_2} P_0(x)Q_0(x^{-1})$$
$$= C_2(x^{-1})P_0(x)Q_0(x^{-1}) = A_2(x). \qquad \square$$

When $K = Z$ and $A_1(x)$, $A_2(x)$ arise from multisets, so that their coefficients are nonnegative integers, it can happen that $A_1(x) = P(x)Q(x)$ and $A_2(x) = P(x)Q(x^{-1})$ where some of the coefficients of $P(x)$ and $Q(x)$ are negative. For instance, the example given at the end of the introduction yields the homometric pair in $Z[R]$

$$A_1(x) = 1 + x + x^5 + x^7 + x^8 + x^{10} + x^{12}$$

and

$$A_2(x) = 1 + x + x^2 + x^5 + x^7 + x^9 + x^{12},$$

and here the corresponding $P(x)$ and $Q(x)$ are respectively

$$P(x) = x^{5/2}(1 + x + x^2 + x^3 + x^4 + x^5 + x^7)$$

and

$$Q(x) = x^{-5/2}(1 - x^3 + x^5).$$

If it should happen that the coefficients in $P(x)$ and $Q(x)$ are nonnegative integers, then we can find multisets A_1, A_2, U, V in R^n corresponding to $A_1(x)$, $A_2(x)$, $P(x)$, $Q(x)$ and we can write $A = U + V$ and $A = U - V$. This explains the rather vague statement at the end of the introduction about "multisets with possibly negative integer multiplicities".

It is clear from the remark after Theorem 2.1 and the purely algebraic nature of the proof of Theorem 2.2 that a similar theorem is true in $K[Z^n]$. The decomposition takes a slightly different form.

THEOREM 2.3. *Two elements $A_1(x)$, $A_2(x)$ in $K[Z^n]$ are homometric if and only if there exist $P(x)$, $Q(x)$ in $K[Z^n]$, a number $c \in K$ of absolute value one, and $v_1, v_2 \in Z^n$ such that $A_1(x) = x^{v_1}P(x)Q(x)$ and $A_2(x) = cx^{v_2}P(x)Q(x^{-1})$.*

It is also easy to extend our theorem to larger collections of pairwise homometric ring elements as follows.

THEOREM 2.4. *The elements $A_1(x), \cdots, A_k(x) \in K[R^n]$ are pairwise homometric if and only if there exist $P_1(x), \cdots, P_r(x) \in K[R^n]$, $I_1, \cdots, I_k \subseteq \{1, \cdots, r\}$, constants $c_1, \cdots, c_k \in K$ of absolute value one, and vectors $v_1, \cdots, v_k \in R^n$, such that for each j, $1 \leqq j \leqq k$,*

$$A_j(x) = c_j x^{v_j} \prod \{P_i(x): i \in I_j\} \prod \{P_i(x^{-1}): i \in \{1, \cdots, r\} \backslash I_j\}.$$

Proof. The "if" part is easy, like the corresponding part of the proof of Theorem 2.2. For the converse, let $A_1(x) = P_1(x) \cdots P_r(x)$ be a prime factorization of $A_1(x)$ in $K[R^n]$. For $2 \leqq j \leqq k$, it can also be arranged that there exist $I_j \in K$, $c_j\bar{c}_j = 1$, and $v_j \in R^n$ such that

$$A_j(x) = c_j x^{v_j} \prod \{P_i(x): i \in I_j\} \prod \{P_i(x^{-1}): i \in \{1, \cdots, r\} \backslash I_j\}.$$

Let $I_1 = \{1, \cdots, r\}$, $c_1 = 1$, and $v_1 = 0$ to get the theorem. □

Remark. Theorem 2.4 is also true if Z^n replaces R^n throughout. The terms x^{v_i} do not appear in Theorem 2.2 because they can be absorbed in $P(x)$, $Q(x)$. But in the more general form of Theorem 2.4, they are needed.

We also have what may seem a rather surprising result from this presentation. Let us say that an element $A(x) \in K[R^n]$ is *symmetric* if there exists $v \in R^n$ such that $A(x^{-1}) = x^v A(x)$. We will say that $A(x)$ is *semi-symmetric* if there exists $v \in R^n$ and $c \in K$ of absolute value one such that $A(x^{-1}) = cx^v A(x)$. For multisets $A \subseteq R^n$, A is centrally symmetric about some point of R^n if and only if $A(x)$ is symmetric; also, A is centrally anti-symmetric about some point of R^n if and only if $A(x)$ is semi-symmetric with the constant $c = -1$. The constant c above is called the *constant of symmetry.*

THEOREM 2.5. *If $A_1(x), A_2(x) \in K[R^n]$ are both semi-symmetric and $A_1(x)$ is homometric to $A_2(x)$, then there exist $v \in R^n$ and $c \in K$ of absolute value one such that $A_1(x) = cx^v A_2(x)$. Moreover, if $A_1(x)$ and $A_2(x)$ have the same constants of symmetry, then c can be taken in $\{-1, 1\}$.*

Proof. We know that $A_i(x^{-1}) = c_i x^{v_i} A_i(x)$ for some $v_i \in R^n$ and $c_i \in K$ of absolute value one, for $i = 1, 2$. Since $A_1(x)A_1(x^{-1}) = A_2(x)A_2(x^{-1})$, we have $c_1 x^{v_1} A_1(x)^2 = c_2 x^{v_2} A_2(x)^2$. Let $c \in K$ of absolute value one satisfy $c^2 = c_1 \bar{c}_2$. Then taking square roots, $A_1(x) = \varepsilon c x^v A_2(x)$ where $\varepsilon \in \{-1, 1\}$ and $v = (v_1 - v_2)/2$. If $c_1 = c_2$, then c can be taken to be one. □

COROLLARY 2.6. *If $A_1(x), A_2(x) \in K[R^n]$ are homometric and both are symmetric (or anti-symmetric), then $A_1(x) = \varepsilon x^v A_2(x)$ for some $\varepsilon \in \{-1, 1\}$ and $v \in R^n$.*

Remark. Theorem 2.5 tells us that a centrally symmetric multiset $A \subseteq R^n$ can be essentially reconstructed from ΔA. In the more general context of homometry discussed in § 3, a theorem which is more general than Theorem 2.5 can be obtained using complex analysis; the proof above though is very simple and purely algebraic.

Let us say that $A(x) \in K[R^n]$ is *reconstructible* if whenever $B(x) \in K[R^n]$ is homometric to $A(x)$, we have $B(x) = cx^v A(x)$ or $B(x) = cx^v A(x^{-1})$ for some $v \in R^n$ and $c \in K$ of absolute value one. In view of Theorem 2.5, being reconstructible is related to being symmetric; but the two concepts are not the same, as the following theorem shows.

THEOREM 2.7. *An element $A(x) \in K[R^n]$ is reconstructible if and only if $A(x)$ has at most one prime factor counting multiplicities which is not semi-symmetric.*

Proof. First, suppose $A(x)$ is not reconstructible. Then there exists $B(x)$ homometric to $A(x)$ such that $B(x) \neq cx^v A(x^\varepsilon)$ for all $v \in R^n$, $\varepsilon \in \{-1, 1\}$, and $c \in K$ of absolute value one. By Theorem 2.2, there exist $P(x)$, $Q(x) \in K[R^n]$ and $c \in K$ of absolute value one such that $A(x) = P(x)Q(x)$ and $B(x) = cP(x)Q(x^{-1})$. If $Q(x)$ is semi-symmetric, then $B(x) = c_1 x^v A(x)$ for some $v \in R^n$ and $c_1 \in K$ of absolute value one. So, $Q(x)$ is not semi-symmetric. By a similar argument, $P(x)$ is not semi-symmetric. Therefore, each of $P(x)$ and $Q(x)$ have prime factors (possibly the same prime in $K[R^n]$) which are not semi-symmetric. So if $A(x)$ is not reconstructible, then there exist two prime factors $P_1(x)$ and $P_2(x)$ of $A(x)$ (with $P_1^2(x)$ being a factor of $A(x)$ if $P_1(x) = P_2(x)$) such that $P_1(x)$ and $P_2(x)$ are not semi-symmetric.

Conversely, suppose $A(x)$ is reconstructible, but it has two prime factors counting multiplicities which are not symmetric, say $P_1(x)$ and $P_2(x)$. Let $A(x) = P_1(x)P_2(x)Q(x)$. For convenience of notation, if $B_1(x), B_2(x) \in K[R^n]$, then $B_1(x) \sim B_2(x)$ will mean that $B_1(x) = cx^v B_2(x)$ for some $v \in R^n$ and some $c \in K$ of absolute value one. Since $A(x)$ is reconstructible, if $B(x)$ is homometric to $A(x)$, then $B(x) \sim A(x)$ or $B(x) \sim A(x^{-1})$. Hence, $A(x) \sim P_1(x^{-1})P_2(x)Q(x)$ or $A(x) \sim P_1(x)P_2(x^{-1})Q(x^{-1})$. In the former case, $P_1(x)P_2(x)Q(x) \sim P_1(x^{-1})P_2(x)Q(x)$ and so

$P_1(x) \sim P_1(x^{-1})$, contradicting the assumption that $P_1(x)$ is not semi-symmetric. Hence, we have the latter case and so

$$P_2(x)Q(x) \sim P_2(x^{-1})Q(x^{-1}). \tag{2}$$

A similar argument with $P_1(x)P_2(x^{-1})Q(x)$ replacing $P_1(x^{-1})P_2(x)Q(x)$ shows

$$P_1(x)Q(x) \sim P_1(x^{-1})Q(x^{-1}). \tag{3}$$

Since $P_1(x)$ is not semi-symmetric, $Q(x) \not\sim Q(x^{-1})$ by (2). Hence, $A(x) = P_1(x)P_2(x)Q(x) \not\sim P_1(x)P_2(x)Q(x^{-1})$. Since $A(x)$ is reconstructible, we must have $P_1(x)P_2(x)Q(x) \sim P_1(x^{-1})P_2(x^{-1})Q(x)$. That is,

$$P_1(x)P_2(x) \sim P_1(x^{-1})P_2(x^{-1}). \tag{4}$$

This gives us, using (2), (3) and (4),

$$P_1(x)^2P_2(x)Q(x) \sim P_1(x)P_1(x^{-1})P_2(x^{-1})Q(x) \sim P_1(x^{-1})^2P_2(x^{-1})Q(x^{-1})$$
$$\sim P_1(x^{-1})^2P_2(x)Q(x).$$

Hence, $P_1(x)^2 \sim P_1(x^{-1})^2$. Taking square roots gives $P_1(x) \sim P_1(x^{-1})$, a contradiction. □

Remark. Theorem 2.7 shows how to distinguish symmetry or semi-symmetry and reconstructibility. For instance, $A_1(x) = (1+x+x^3)(1+x^2+x^3)$ is symmetric but not reconstructible in $Z[R]$. Also, $A_2(x) = (1+x+x^3)(1+x^2+x^3)(1-x)$ is semi-symmetric, not symmetric, and not reconstructible in $Z[R]$. On the other hand, $A_3(x) = (1+x+x^3)(1+x^4+x^8)$ is reconstructible in $Z[R]$ but it is not semi-symmetric. Notice also that the coefficient ring K can have a great effect on how strong a restriction it is for $A(x)$ to be reconstructible. For instance, let $A(x) \in \mathbb{C}[Z]$ and choose $v \in Z$ such that $x^v A(x) = P(x)$ is a polynomial in x. Then $A(x)$ is reconstructible in $\mathbb{C}[Z]$ if and only if at most one of the roots of $P(x)$ in $\mathbb{C}$ does not have absolute value one.

3. Applications. The structure theorems of § 2 provide an entirely algebraic approach to one of the phase retrieval problems arising in diffraction by distributions which consist of a finite number of atoms. There are many texts which discuss the theory of Fresnel and Fraunhofer diffraction (the latter being the limiting case of the general Fresnel diffraction and is the one to which the theory of homometric sets is most applicable). (See [2], [3], [4].) The general idea is that a bounded point distribution in space (say R^2 or R^3) will diffract electromagnetic radiation, whose wave fronts are essentially planar throughout the region of diffraction, in such a manner that, at large distances from the diffracting atoms and near the axis of diffraction, the diffraction has a complex amplitude in space which is given by the Fourier transform of the original point distribution. That is, if the original distribution A is a finite sum $\sum a_x \delta_x$ where a_x is a positive integer associated with the Dirac mass δ_x at $x \in R^n$, $n = 2, 3$, then the diffracted wavefront in Fraunhofer diffraction has a complex amplitude given by $\hat{A}(y) = \sum a_x \exp(-ix \circ y)$ where $i = \sqrt{-1}$ and $x \circ y$ is the usual scalar product in R^n. The whole numbers a_x correspond to the "size" of the Dirac point masses δ_x under diffraction; for example, in X-ray diffraction this usually is proportional to the number of electrons in the atom at site $x \in R^n$. This is of course the ideal situation; in reality there will be many sources of scattering which cause a_x to be a Gaussian variable in R^n centered at x with maximum amplitude given by a_x. Also, there are many difficulties in measuring $\hat{A}$ because one needs to sample the diffracted wavefront at different points in R^n or rotate the plane of the incoming wavefront before

diffraction. But once $\hat{A}$ is determined up to an experimental error, then A is completely determined up to corresponding errors by the Fourier inversion theorem.

The phase retrieval problem arises because in an actual experimental situation it is the absolute value $|\hat{A}(y)|$ of $\hat{A}(y)$ which is measured when taking photographs of the diffracted wavefront or when using a Geiger counter to measure the intensity (i.e., the absolute amplitude) of $\hat{A}$ in space. With some computer assisted computations, one can usually determine $|\hat{A}|$ fairly well in a given experimental situation. The phase retrieval problem is that the phase of $\hat{A}$ is not given by $|\hat{A}|$ and $|\hat{A}|$ does not uniquely determine $\hat{A}$ or A. However, given $|\hat{A}|$ or $|\hat{A}|^2$, one has $\hat{A}(y)\overline{\hat{A}(y)}$ determined. The *Patterson function of* A (or the *auto-correlation function*, as it is also called) is the point distribution $A * A^*$ where A^* is the usual *involution* given by $A^*(x) = \sum \bar{a}_x \delta_{-x}$ and $A * A^*$ is the *convolution* of A and A^*. That is, $A * A^* = \sum c_z \delta_z$ where $c_z = \sum \{a_x \bar{a}_y : x - y = z\}$ for all $z \in R^n$. It is easy to check that the Patterson function $A * A^*$ has a Fourier transform equal to $\hat{A}(y)\overline{\hat{A}(y)}$ since $\hat{A}^*(y) = \overline{\hat{A}(y)}$ for all $y \in R^n$. Hence, it is the Patterson function $A * A^*$ which is determined by experimental measurement in Fraunhofer diffraction.

We see then that two point distributions $A = \sum a_x \delta_x$ and $B = \sum b_x \delta_x$ are homometric if and only if they appear the same in diffraction experiments that measure only $|\hat{A}|$ and $|\hat{B}|$. This shows how the structure theorems in § 2 are directly applicable to the phase retrieval problem of determining A from a knowledge of $\hat{A}$, that is, of $A * A^*$, only. One could, throughout § 2, have used the group ring notation or the very similar notation of Dirac masses and convolution in R^n. In fact, all the theorems on homometry apply to the ring under pointwise multiplication of exponential polynomials of the form $P(y) = \sum a_x \exp(-ix \circ y)$ where $a_x \in K$ is nonzero only finitely many times. The theorems that were proved for this ring in § 2, like Theorem 2.2, bear comparing with the theorems of Ritt [5], [6], [7] that deal with more general exponential polynomials of one variable. Nonetheless, we have used polynomial notation throughout § 2 because of its general familiarity and because of the availability of good computer programs for factoring polynomials. Here is the general algebraic approach to phase retrieval as it might be applied in an experimental context.

Step 1. Through experimental measurement, the Patterson function $A * A^*$ is determined. Notice that $A * A^*$ is symmetric. Say $A * A^* = \sum p_v \delta_v$ with the coefficients $p_v \in K$.

Step 2. Let $V = \{v \in R^n : p_v \neq 0\}$. Then V is a finite set and spans an m-dimensional lattice $L = \{\sum_{j=1}^{l} c_j v_j : c_i \in Z,\ v_i \in V,\ i = 1, \cdots, l\}$. A basis for L is found, say some $e_1, \cdots, e_m \in L$. This determination is greatly simplified if the elements of V are already at points of the natural lattice $Z^n \subseteq R^n$ where R^n is the underlying space in which A is given, but in general L may be a higher dimensional lattice. At least m is no larger than the number of elements in V.

Step 3. Express each $v \in V$ as $v = \sum_{j=1}^{m} v_j e_j$ where $v_1, \cdots, v_m \in Z$. Then let $P(x)$ be the element of $K[Z^m]$ given by $P(x) = \sum p_v x^v$. By multiplying $P(x)$ by some suitable x^w with $w \in L$ (that is, by translation the Patterson function $A * A^*$), we get an element $Q(x) = x^w P(x)$ in $K[Z^m]$ which is a polynomial in positive powers of $x_1, \cdots, x_m$.

Step 4. Using some algebraic factoring technique and/or a computer program like IBM's Scratchpad, we now factor $Q(x)$ in the polynomial ring $K[x_1, \cdots, x_m]$ into its prime factors. Because $Q(x)$ is essentially the Patterson function $A * A^*$ in a polynomial form, there must be primes $Q_1(x), \cdots, Q_r(x) \in K[x_1, \cdots, x_m]$ and a vector $w_0 \in L$ such that $Q(x) = x^{w_0} \prod_{j=1}^{r} Q_j(x) \prod_{j=1}^{r} Q_j(x^{-1})$. We then form all possible products $A_I(x)$ of the form $A_I(x) = \prod \{Q_j(x) : j \in I\} \prod \{Q_j(x^{-1}) : j \in \{1, \cdots, r\} \backslash I\}$ where $I \subseteq \{1, \cdots, r\}$ as in Theorem 2.4. These products $A_I(x)$ each have the form $\sum \{a_v(I) x^v : v \in$

$L\}$, where the coefficients $a_v(I) \in K$, and depend on I, but are nonzero only finitely many times. We use the notation $A_I = \sum a_v(I)\delta_v$. For each I, A_I is a point distribution in $L \subseteq R^n$ which is homometric to A. By Theorems 2.2 and 2.4, up to translations and multiplications by $c \in K$ of absolute value one, all possible point distributions which are homometric with A must appear in the set $\{A_I : I \subset \{1, \cdots, r\}\}$. Of course, in particular $A \sim A_I$ for some I (not necessarily $I = \{1, \cdots, r\}$). In this way, using just the Patterson function $A * A^*$, we can rediscover A and all point distributions with the same Patterson function. There may be as many as 2^r distinct distributions A_I in general; so the list of the A_I may take a good deal of computer time to enumerate. This procedure can be considerably shortened by just choosing not to reflect any $Q_i(x)$, $i = 1, \cdots, r$, which is semi-symmetric. Also, if only A_I with positive coefficients could possibly be A, then the other A_I can be ignored in a specific search for A.

The program for phase retrieval outlined in Steps 1–4 above is the first complete algebraic procedure for recovering a general bounded point distribution from its Patterson function alone. This procedure requires no additional information about A (e.g., feasibility as a chemical molecule undergoing X-ray diffraction), although this information can certainly be useful in locating feasible distributions among the A_I. When starting with A which has only positive integral coefficients, the list of homometric distributions A_I may include many which have nonpositive coefficients. There are many possible experimental reasons for a coefficient a_x to include a phase shift and be a general complex value. This occurs typically in neutron diffraction and also in X-ray diffraction where there is a phase shift in the diffracting wavefront due to absorption at the atom site $x \in R^n$. (See [3] for a discussion of this phenomenon.) This means that we generally want to include all A_I in our list of feasible distributions, even those with nonpositive coefficients.

In the method outlined in Steps 1–4, it was assumed at the outset that the Patterson function is known precisely. In experimental measurements, there will always be some errors to take into account. In factoring polynomials, especially in $Z[R^n]$, errors in the coefficients or exponents can generally have an enormous effect on the factorization. This is analogous to how the set of solutions of a system of linear equations can be changed dramatically (from an infinite set to a singleton, for instance) by very small changes in the coefficients. There are several possible methods of dealing with these errors. First, if the Patterson function is known to have only integer coefficients at the points of some known lattice Z^m, $Z^m \subset R^n$, then the experimental errors can be made small enough to identify the Patterson function precisely. Second, when possible we can choose the coefficient ring to be $\mathbb{C}$ in order to improve the chances of factorization occurring; this works especially well in $\mathbb{C}[Z]$. Then one would multiply these factors together to get approximate factorizations in the smaller ring $Z[Z]$. Finally, one should be prepared to incorporate in any computer program some random sampling of many Patterson functions within the allowed experimental deviation from the unknown actual Patterson function. The only a priori restriction on these sample Patterson functions is symmetry. Each of these sample Patterson functions is then used for the entire algebraic procedure Steps 1–4. It is advisable when doing this sampling to try to make the dimension of the lattice L in Step 2 as small as possible by appropriate choice of the atom sites within the allowed errors. This will increase the number of prime factors in general (if there is any chance that A has many prime factors) and it will make the factorization easier to accomplish. There are probably many difficulties that errors in measurement will create for our algebraic program; but there should also be many methods, for example the ones mentioned above, to counteract these difficulties.

Unfortunately, the algebraic techniques above only apply to finite distributions. If one is dealing with periodic point distributions which are effectively (up to the accuracy of experimental measurement) spread throughout space (as is the case in X-ray crystallography for instance), then the structure theorems of § 2 do not apply. See [1] for some examples and a use of the convolution method of § 2 to provide large families of nonisometric, mutually homometric crystal structures. One can view the failure of the algebraic technique as a result of having to replace $K[R^n]$ by a group ring $K[G]$ where G is some finite cyclic group Z_n or $Z_n \times Z_n$ or $Z_n \times Z_n \times Z_n$. Such a group ring has zero divisors and is not locally a unique factorization domain. The reason for the change in the group ring is that for Fraunhofer diffraction of crystals, it is the absolute value of a Fourier series which is measured experimentally and all vectors in the support of the Patterson function have to be identified modulo the crystal's lattice structure. So there is still no adequate structure theory for phase retrieval in Fraunhofer diffraction of periodic distributions.

Another direction in which one could hope to extend our technique is diffraction of continuous distributions with compact support. Here Theorem 2.2 fails to hold except as a method of providing large classes of nonisometric, mutually homometric continuous distributions. However, Theorem 2.5 does hold in this generality, as one can see using an argument in several complex variables as applied to the extension of the Fourier transform to the complex domain $\mathbb{C}^n$, as in the Paley–Wiener theorem. But as with periodic distributions, there is no adequate structure theory yet for phase retrieval in Fraunhofer diffraction of continuous distributions.

Acknowledgment. The authors wish to thank David Yun for some helpful correspondence concerning the capabilities of IBM's Scratchpad program.

REFERENCES

[1] J. BERMAN AND J. ROSENBLATT, *The characterization of homometric structures*, preprint.
[2] M. BORN AND E. WOLF, *Principles of Optics*, Pergamon Press, Oxford, 1970.
[3] J. M. COWLEY, *Diffraction Physics*, North-Holland, Amsterdam, 1975.
[4] R. HOSEMANN AND S. N. BAGCHI, *Direct Analysis of Diffraction by Matter*, North-Holland, Amsterdam, 1962.
[5] J. F. RITT, *A factorization theory for functions*, Trans. Amer. Math. Soc., 29 (1927), pp. 584–596.
[6] ———, *Algebraic combinations of exponentials*, Trans. Amer. Math. Soc., 31 (1929), pp. 654–679.
[7] ———, *Zeros of exponential polynomials*, Trans. Amer. Math. Soc., 31 (1929), pp. 680–686.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 3, September 1982

0196-5212/82/0303-0010 $01.00/0

THE COMPLEXITY OF THE PARTIAL ORDER DIMENSION PROBLEM*

MIHALIS YANNAKAKIS†

Abstract. The dimension of a partial order P is the minimum number of linear orders whose intersection is P. There are efficient algorithms to test if a partial order has dimension 1 or 2. We prove that it is NP-complete to determine if a partial order has dimension 3. As a consequence, several other related dimension-type problems are shown to be NP-complete.

Key words. partial order, dimension, NP-complete, interval dimension, threshold dimension, boxicity, cubicity

1. Introduction. A *partial order* P of a finite[1] set N is an irreflexive, transitive binary relation on N; i.e., $(x, x) \notin P$ for each $x \in N$, and if (x, y) and $(y, z) \in P$ then $(x, z) \in P$. A *linear order* L of N is a partial order which contains (x, y) or (y, x) for any two distinct elements x, y of N. The linear order L is a (*linear*) *extension* of a partial order P if $P \subseteq L$. A partial order P can be viewed as a transitive directed acyclic graph (DAG) with set of nodes N and arcs $x \to y$ for $(x, y) \in P$. A linear order L is then a complete DAG; it is a linear extension of P if P is a subgraph of L.

The intersection of any set of partial orders of a set N is obviously also a partial order. The *dimension* $d(P)$ of a partial order P of N is the minimum number of linear orders whose intersection is P [DM]. It is a well-defined parameter: every partial order is the intersection of some linear extensions of it. Moreover, $d(P) \leqq |N|/2$[H]. A geometric interpretation of the dimension (and justification of the term) is the following. Let π be a mapping from N to distinct points of the d-dimensional Euclidean space E^d. Let $P(\pi)$ be the partial order of N defined by: $(x, y) \in P$ if and only if each coordinate of $\pi(x)$ is less than the corresponding coordinate of $\pi(y)$. The dimension of a partial order P of N is the minimum d for which there exists such a mapping π from N to E^d with $P(\pi) = P$ [O].

Clearly, a partial order has dimension 1 if and only if it is a linear order. Duschnik and Miller [DM] proved a necessary and sufficient condition for a partial order P to have dimension 2. Two elements x and y of N are *comparable* if (x, y) or (y, x) belongs to P; otherwise they are *incomparable*. The *incomparability graph* of P is an undirected graph $I(P)$ with N as its set of nodes and edges connecting the pairs of incomparable elements. [DM] proved that P has dimension 2 if and only if $I(P)$ is *transitively orientable*; i.e., the edges of $I(P)$ can be oriented so that the resulting directed graph is transitive. (Such a graph is sometimes called a *comparability graph*.) This condition combined with an efficient algorithm for the recognition of transitively orientable graphs [PLE] gives a polynomial algorithm to test if a partial order has dimension at most 2. A complete set of forbidden subgraphs of such partial orders is given in [K], [TM].

In this paper we will prove that it is NP-complete to determine if the dimension of a partial order is at most 3, and consequently the same holds also for any fixed $k \geqq 3$. The complexity of the partial order dimension problem was unknown for any fixed $k \geqq 3$ and for an arbitrary k; it is one of the open problems in the list of Garey and Johnson [GJ]. For an exposition on NP-completeness see [GJ].

* Received by the editors July 8, 1981 and in revised form September 3, 1981.
† Bell Laboratories, Murray Hill, New Jersey 07974.
[1] All sets in this paper will be finite.

2. Preliminaries. Before describing the reduction, let us get some insight into the problem. Let P be a partial order of N (transitive DAG on N) and suppose that N is partitioned into two sets S, S' so that there is no arc of P directed from a node of S' to a node of S. Let $B(P)$ be the bipartite graph with set of nodes N and set of edges $\{[x, y] | x \in S, y \in S', x \text{ and } y \text{ incomparable}\}$. Consider now a linear extension L of P, and let $\bar{L}$ be the bipartite graph with set of nodes N and set of edges $\{[x, y] | x \in S, y \in S', (y, x) \in L\}$; i.e., $\bar{L}$ contains those arcs (without the direction) which are directed from nodes of S' to nodes of S. Since P has no such arcs and L is a linear extension of P, $\bar{L}$ must be a subgraph of $B(P)$.

We say that two edges $[x, y]$, $[z, w]$, of a graph are *independent* if the nodes x, y, z, w are distinct and the subgraph induced by them consists of exactly these two edges. Suppose that $\bar{L}$ had two independent edges $[x, y]$, $[z, w]$ with $x, z \in S$ and $y, w \in S'$. From the definition of $\bar{L}$ then, L would contain (y, x), (w, z), (z, y) and (x, w); i.e., L would contain a cycle $y \to x \to w \to z \to y$, contradicting the fact that L is a linear order (complete DAG). Thus, $\bar{L}$ has no pair of independent edges. We call a bipartite graph with this property, a *chain graph*. It is also characterized by the property that the neighborhoods Γ_x (sets of nodes adjacent to x) of nodes x in S are totally ordered by set inclusion; i.e. for every x, y in S, either $\Gamma_x \subseteq \Gamma_y$ or $\Gamma_y \subseteq \Gamma_x$ [Y].

Suppose now that P has dimension d, and let $L_1, \cdots, L_d$ be linear extensions of P with intersection P. Let $\bar{L}_1, \cdots, \bar{L}_d$ be the bipartite graphs that we defined above. Each $\bar{L}_i$ is a chain subgraph of $B(P)$. Since the intersection of the L_i's is P, for every edge $[x, y]$ of $B(P)$ with $x \in S$, $y \in S'$, the arc (y, x) must appear in at least one of the L_i's; thus $[x, y]$ is covered by (appears in) at least one of the $\bar{L}_i$'s. For a bipartite graph G, let $\text{ch}(G)$ be the minimum number of chain subgraphs of G that cover all the edges of G. We have shown:

LEMMA 1. $d(P) \geqq \text{ch}(B(P))$.

Thus, for example, if P is the *crown* on nodes $\{v_1, \cdots, v_k, v'_1, \cdots, v'_k\}$ with arcs $v_i \to v'_j$ for $i \neq j$, and we take $S = \{v_1, \cdots, v_k\}$, $S' = \{v'_1, \cdots, v'_k\}$, then $B(P)$ consists of k pairwise independent edges $[v_1, v'_1], \cdots, [v_k, v'_k]$, and $d(P) \geqq \text{ch}(B(P)) = k$.

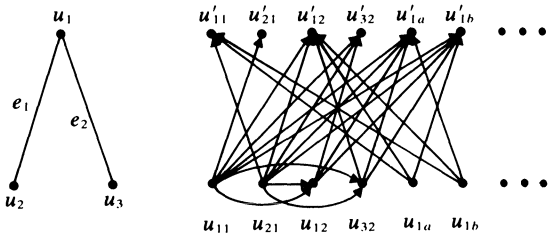
3. The reduction. The reduction is from the chromatic number 3 problem; i.e., given a graph G determine if the nodes of G can be colored with 3 colors so that adjacent nodes receive different colors. This problem was shown NP-complete in [GJS]. From G we will construct a partial order P so that G can be colored with 3 colors if and only if $d(P) \leqq 3$.

Let $V = \{u_1, \cdots, u_n\}$ be the nodes of G, and $E = \{e_1, \cdots, e_m\}$ its edges. P is a partial order on the union N of two disjoint sets S and S'. S contains two nodes u_{ia}, u_{ib} for every node u_i of V, and two nodes u_{ik}, u_{jk} for every edge $e_k = [u_i, u_j]$ in E. S' contains the primed versions of the nodes in S; thus, altogether N has $4n + 4m$ elements. The partial order P is defined as follows.

$$
\begin{aligned}
P = \{(u_{ia}, u'_{it}) | 1 \leqq i \leqq n, t \neq a\} &\cup \{(u_{ib}, u'_{it}) | 1 \leqq i \leqq n, t \neq b\} \\
&\cup \{(u_{ik}, u'_{jl}) | 1 \leqq i, j \leqq n, 1 \leqq k \leqq m, l > k \text{ or } l = a \text{ or } b\} \\
&\cup \{(u_{ik}, u'_{jk}) 1 \leqq k \leqq m, e_k = [u_i, u_j]\} \\
&\cup \{(u_{ik}, u_{jl}) | 1 \leqq i, j \leqq n, 1 \leqq k < l \leqq m\}.
\end{aligned}
$$

Notice that all arcs between S and S' are directed from S to S'. Let $B(P)$ be the bipartite graph defined in the previous section. Let Q be the nodes u_{ik} in S that correspond to edges of G (i.e., with $1 \leqq k \leqq m$), R the rest of the nodes in S (i.e. with $k = a, b$), and let Q', R' be the analogous subsets of S'. In simple words, $B(P)$ has

the following structure: a node in Q is connected to its primed version and to all nodes of Q' with a strictly smaller second index; a node in R is connected to its primed version and all nodes of S' with a different first index. In Fig. 1 we give an example of the construction; only part of the partial order is shown for clarity.



(a) A graph G. (b) Part of the partial order P.

FIG. 1

LEMMA 2. *If* ch $(B(P)) \leqq 3$ *then* G *can be colored with* $\leqq 3$ *colors.*

Proof. Suppose that ch $(B(P)) \leqq 3$, and let B_1, B_2, B_3 be three chain subgraphs of $B(P)$ that cover it. Consider the subgraph H_i of $B(P)$ induced by all nodes u_{it}, u'_{it} with first index i. It has three connected components: the edge $[u_{ia}, u'_{ia}]$, the edge $[u_{ib}, u'_{ib}]$ and the subgraph induced by the u_{ik}, u'_{ik} with $1 \leqq k \leqq m$. Since a chain graph cannot contain two independent edges, none of the B_j's can contain two edges from different components of H_i. Since H_i has three components, all edges of the third component are in the same B_j; color node u_i with the index of this B_j.

We must show now that this is a legal coloring of G. Suppose that there are two adjacent nodes u_i, u_j with the same color, say color 1. Let $e_k = [u_i, u_j]$. From the definition of the coloring we have $[u_{ik}, u'_{ik}]$, $[u_{jk}, u'_{jk}] \in B_1$. But these two edges are independent in $B(P)$, hence also in B_1, contradicting the assumption that B_1 is a chain graph. □

LEMMA 3. *If* G *can be colored with* $\leqq 3$ *colors then* $d(P) \leqq 3$.

Proof. Suppose that G can be colored with 3 colors, and let C_1, C_2, C_3 be the sets of nodes that receive color 1, 2, 3 respectively in a legal coloring of G. We will construct three linear orders L_1, L_2, L_3 whose intersection is P. We will show only L_1 corresponding to the color class C_1; the two other linear orders are analogous.

We need some notation for describing linear orders. We shall write a linear order as a string where every element is less than the elements to its right. If X is a set then X will stand also for an arbitrary linear order of X. If F_1, F_2 are linear orders of disjoint sets X_1, X_2, then F_1F_2 is the concatenation of the two strings. If $I = \{i_1, \cdots, i_k\}$ is an index set with $i_1 < i_2 < \cdots < i_k$, and $F_{i_1}, \cdots, F_{i_k}$ are linear orders of disjoint sets $X_{i_1}, \cdots, X_{i_k}$, then we will denote $F_{i_1}F_{i_2} \cdots F_{i_k}$ by $\langle F_i \uparrow i \in I \rangle$, and $F_{i_k}F_{i_{k-1}} \cdots F_{i_1}$ by $\langle F_i \downarrow i \in I \rangle$.

Let $R_1 = \{u_{ia}, u_{ib} | u_i \in C_1\}$ and similarly for R'_1, R_2 etc. Let $e_k = [u_i, u_j]$ be an edge of G. If none of u_i, u_j has color 1 then $E_k = \{u_{ik}, u_{jk}\}$. If one of e_k's nodes, say u_i, has color 1 then E_k is the order $u_{jk}u'_{ik}u_{ik}$.

Let u_i be a node that receives color 2 or 3. Define a linear order K_i on $\{u_{ia}, u_{ib}, u'_{ia}, u'_{ib}\} \cup \{u'_{ik} | u_i \in e_k\}$ as follows. If u_i has color 2 then K_i is $u_{ib}u'_{ia}u_{ia}u'_{ib}$ followed by the elements of the second set in decreasing order of the second index (k). If u_i has color 3 then K_i is $u_{ia}u'_{ib}u_{ib}u'_{ia}$ followed by the elements of the second set again in decreasing order of k.

The linear order L_1 now is $R_1 < E_k \uparrow k \in \{1, \cdots, m\} > R_1' \langle K_i \uparrow u_i \in C_2 \rangle \langle K_i \downarrow u_i \in C_3 \rangle$. The two other orders are defined in a cyclically symmetric fashion.

It is straightforward to verify that each L_i is a linear extension of P. We will show now that for every pair of incomparable elements x, y of P, (x, y) is in one of the L_i's.

Case 1. $x \in S'$, $y \in S$. This case amounts to showing that the chain subgraphs $\bar{L}_1$, $\bar{L}_2$, $\bar{L}_3$ of $B(P)$ cover all of its edges. Let $a = u'_{ik}$, $y = u_{jt}$, and suppose u_i has color 1, and u_j color c. If $c \neq 1$, or $t \in \{1, \cdots, m\}$ then $(x, y) \in L_1$. Assume then $c = 1$ and $t \in \{a, b\}$. If $i > j$ then $(x, y) \in L_2$; if $i < j$ then $(x, y) \in L_3$. If $i = j$ we must have $k = t$ since x and y are incomparable and $t = a$ or b; then $(x, y) \in L_2$ or L_3.

Case 2. $x, y \in S'$. Let $x = u'_{ik}$, $y = u'_{jt}$. If $i = j$ and $1 \leqq t < k \leqq m$, then, $(x, y) \in L_c$ where c is not the color of u_i. In all other cases there is a node z of S incomparable with x and such that $(z, y) \in P$: If $i \neq j$ or $[i = j$ and $k \in \{a, b\}$, say $k = a]$ then u_{ja} is such a node z. *If* $i = j$ and $k \in \{1, \cdots, m\}$ then we must have $t \in \{a, b\}$ or $t > k$, and u_{ik} is such a node z. From Case 1, (x, z) is in one of the linear extensions L_r of P and therefore $(x, y) \in L_r$ by transitivity.

Case 3. $x \in S$, $y \in S'$. Clearly, there is a $z \in S'$ such that $z \neq y$, $(x, z) \in P$. Since z and y are incomparable, (z, y) is in some L_r from Case 2 and by transitivity $(x, y) \in L_r$.

Case 4. $x, y \in S$. If there is a z in S' incomparable to y with $(x, z) \in P$, then (x, y) is in some L_r as before. If there is no such z, then it is easy to see that $x = u_{ik}$ for some $k \in \{a, b\}$ and $y = u_{jt}$ for some $t \in \{1, \cdots, m\}$. Then $(x, y) \in L_c$ where c is the color of u_i. □

THEOREM 1. *It is* NP*-complete to determine if the dimension of a given partial order is at most* 3.

Proof. Follows immediately from Lemmas 1, 2, 3. □

COROLLARY 1. *It is* NP*-complete to determine if a given bipartite graph can be covered by* 3 *chain subgraphs.*

Proof. Same as for Theorem 1. □

We should mention here that it is possible to determine in polynomial time if a bipartite graph can be covered by 2 chain subgraphs; this follows from the results of Ibaraki and Peled in [IP] and Lemma 7 in the next section.

The *height* of a partial order P is the length of the longest path in (the DAG) P. [Ki] showed that a partial order P can be transformed efficiently into another partial order P' of height 1 with $d(P) \leqq d(P') \leqq d(P) + 1$.[2] Therefore, an efficient algorithm for computing the dimension of partial orders of height 1 would give a good approximation of the dimension of an arbitrary partial order.

COROLLARY 2. *It is* NP*-complete to determine if the dimension of a partial order of height* 1 *is at most* 4.

Proof. Let B_1 be a bipartite graph with S_1, S_1' a bipartition of its nodes. Let B be obtained from B_1 by adding two new nodes u, u' and an edge $[u, u']$. Let $S = S_1 \cup \{u\}$ and $S' = S_1' \cup \{u'\}$. Let P be the partial order of height one with $(x, y) \in P$ if and only if $x \in S$, $y \in S'$ and $[x, y] \notin B$. Notice that B is the graph $B(P)$ that we defined in § 2. We claim that ch $(B_1) \leqq 3$ if and only if $d(P) \leqq 4$ (if and only if ch $(B) \leqq 4$).

(*if*) Suppose that $d(P) \leqq 4$. From Lemma 1 then ch $(B) \leqq 4$. Let G_1, G_2, G_3, G_4 be chain subgraphs of B that cover its edges and suppose without loss of generality that $[u, u'] \in G_1$. Since $[u, u']$ is independent from all the other edges of B, the rest of the edges must all appear in the other three chain subgraphs.

[2] The partial order P' of height 1 has the property $d(P) = \text{ch}\,(B(P'))$; that is, the construction of [Ki] is actually a reduction of the dimension problem to the chain covering problem. It is easy to see also that for any order P' of height 1, $d(P') \leqq \text{ch}\,(B(P')) + 1$—a proof is essentially contained in Corollary 2.

(*only if*) Let G_1, G_2, G_3 be three chain subgraphs of B_1 that cover its edges. From any chain graph G_i we can get a linear order L_i on $S_1 \cup S_1'$ such that $\bar{L}_i = G_i$. To see this, recall from § 2 that the neighborhoods in G_i of all nodes of S_1 (and S_1') are totally ordered by set inclusion. From this it follows easily that the nodes in S_1 can be partitioned into sets $R_1, R_2, \cdots, R_k$ and the nodes in S_1' into sets $R_1', R_2', \cdots, R_k'$ so that the neighborhood of each node in R_j is $\cup_{t>j} R_t'$. (R_1' and/or R_k may be empty.) The linear order $L_i = R_k R_k' R_{k-1} R_{k-1}', \cdots, R_1 R_1'$ satisfies then $\bar{L}_i = G_i$ (see Fig. 2).

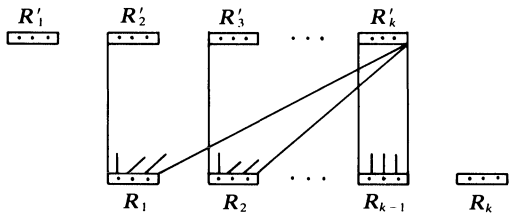


FIG. 2

Let L_1, L_2, L_3 therefore be linear orders with $\bar{L}_i = G_i$. Let F_i be the intersection of the restrictions of L_1, L_2, L_3 on S_1, and F_1' the intersection of the restrictions on S_1'. Let F, F' be any linear orders of S_1, S_1' respectively with $F \cap F_1 = \varnothing$, $F' \cap F_1' = \varnothing$, e.g., F can be the inverse of any topological sort (linear extension) of the DAG F_1.

Let $L_1' = uL_1u'$, $L_2' = uL_2u'$, $L_3' = uL_3u'$, $L_4' = Fu'uF'$, and let P^* be the intersection of the L_i''s. Clearly, all L_i''s are linear extensions of P; thus $P \subseteq P^*$. From our choice of F and F', the restrictions of P^* on S and S' are empty. For $x \in S$, $y \in S'$ with x, y incomparable we have $(x, y) \in L_4'$ or $x = u$, $y = u'$ and $(x, y) \in L_1'$. For $x \in S_1'$, $y \in S_1$ incomparable, (x, y) is in the L_i' that corresponds to the chain subgraph G_i that covers $[x, y]$. Finally, (u', u) is in L_4'. Thus, $P = P^*$. □

COROLLARY 3. *For any $k \geqq 3$ it is* NP*-complete to determine if the dimension of a partial order is at most k.*

Proof. Apply the reduction of Corollary 2, $k-3$ times. □

Note. E. Lawler and O. Vornberger showed recently (and independently) the result in the case of arbitrary dimension; i.e., given partial order P and integer k it is NP-complete to determine if $d(P) \leqq k$ [L].

4. Related problems. In this section we will show that several related dimension-type problems are NP-complete, using Corollary 1.

Interval dimension. Let X be a set of closed intervals on the real line. We can define a partial order P on X, where for x, y in X we have $(x, y) \in P$ if and only if the right endpoint of interval x is to the left of the left endpoint of interval y. A partial order that can be constructed in this way from a set of intervals is called an *interval order.* Clearly, every linear order is an interval order; in this case the intervals can be taken to be distinct points.

The *interval dimension* of a partial order P, denoted $\mathrm{id}(P)$, is the minimum number of interval orders whose intersection is P [TB]. Since every linear order is also an interval order, we have $\mathrm{id}(P) \leqq d(P)$.

Interval orders of height 1 are closely related to chain graphs. A necessary and sufficient condition for a partial order P to be an interval order is that P does not contain a pair of independent arcs, i.e., two arcs (x, y) and (u, v) with x, y, u, v distinct elements and such that the subgraph of P induced by them consists of exactly these

two arcs [F]. Thus, a partial order of height 1 is an interval order if and only if its underlying graph (its comparability graph) is a chain graph.

Let P be a partial order of height 1 and let S be the set of elements of height 1 and S' the set of elements of height 0. Clearly all arcs of P are directed from S to S'. Let $G(P)$ be the underlying graph of P, and $B(P)$ the graph that we defined in § 2; i.e., $B(P)$ is the bipartite graph with the set of edges $\{[x, y] | x \in S, y \in S', (x, y) \notin P\}$. It follows easily from the definitions that $G(P)$ is a chain graph if and only if $B(P)$ is a chain graph.

LEMMA 4. *Let P be a partial order of height* 1. *Then* $\mathrm{id}(P) = \mathrm{ch}(B(P))$.

Proof.

(1) $\mathrm{id}(P) \leqq \mathrm{ch}(B(P))$. Let $B_1, \cdots, B_k$ be chain subgraphs of $B(P)$ that cover its edges. For each B_i define the partial order $P_i = \{(x, y) | x \in S, y \in S', [x, y] \notin B_i\}$. Since B_i is a chain subgraph of $B(P)$, P_i is an interval order that contains P. Since the B_i's cover the edges of $B(P)$, the intersection of the P_i's is P.

(2) $\mathrm{ch}(B(P)) \leqq \mathrm{id}(P)$. Let $P_1, \cdots, P_k$ be interval orders whose intersection is P. For each i, let P_i' be the subgraph of P_i that consists of those arcs of P_i that are directed from S to S'. Let (x, y), (u, v) be two arcs of P_i' with the nodes $x, u \in S$ and $y, v \in S'$ distinct. Since P_i is an interval order, these two arcs cannot be independent in P_i. Suppose that P_i contains an arc from one of $\{x, y\}$ to one of $\{u, v\}$ (the other case is symmetric). Then, by transitivity, (x, v) is in P_i, and therefore also in P_i'. Thus, P_i' is an interval order of height 1, and consequently $B(P_i')$ is a chain graph. Since the intersection of the P_i's is P and all arcs of P are directed from S to S', the intersection of the P_i''s is also P. Therefore, the $B(P_i')$'s are chain subgraphs of $B(P)$ that cover its edges. □

COROLLARY 4. *It is* NP*-complete to determine if the interval dimension of a partial order of height* 1 *is at most* 3.

Proof. Follows from Corollary 1 and Lemma 4. □

[TM] presents a characterization of partial orders of height 1 that have interval dimension 2, in terms of forbidden subgraphs. However, the interval dimension 2 problem for general partial orders is open.

Boxicity. Let X be a set of closed intervals on the real line. We can construct a graph G with the intervals as nodes, and an edge between any two intervals with a nonempty intersection. A graph that can be constructed in this way from a set of intervals is called an *interval graph*. Thus, an interval graph is the incomparability graph of an interval order.

If $G_1, \cdots, G_k$ are graphs with the same set of nodes, their intersection is a graph with the same nodes and with those edges that are contained in all the G_i's. The *boxicity* of a graph G, denoted $b(G)$, is the minimum number of interval graphs whose intersection is G. A geometric interpretation (and justification of the term) is the following. Let X be a set of boxes in the k-dimensional space with sides parallel to the coordinate axis. Their intersection graph has set of nodes X and an edge between any two boxes with a nonempty intersection. The boxicity of a graph G is the minimum k such that G is the intersection graph of a set of such boxes in the k-dimensional space [R]. Thus, G has boxicity 1 if and only if it is an interval graph, boxicity 2 if and only if it is the intersection graph of rectangles in the plane with sides parallel to the axis, etc.

LEMMA 5. *Let $\bar{B}$ be the complement of a bipartite graph B. Then,* $b(\bar{B}) = \mathrm{ch}(B)$.

Proof. At first let us show that the complement $\bar{G}$ of a bipartite graph G is an interval graph if and only if G is a chain graph. If G is a chain graph with S, S' a

bipartition of its nodes, then the partial order P obtained from G by directing all its edges from S to S' is an interval order. Therefore, $\bar{G}$, the incomparability graph of P, is an interval graph. Conversely, if $\bar{G}$ is an interval graph then it is the incomparability graph of an interval order P. Therefore G, the underlying graph of P, does not contain a pair of independent edges. Since G is also bipartite, it is a chain graph.

(1) $b(\bar{B}) \leqq \text{ch}(B)$. Let $B_1, \cdots, B_k$ be chain subgraphs of B that cover its edges. Their complements $\bar{B}_1, \cdots, \bar{B}_k$ are interval graphs whose intersection is $\bar{B}$.

(2) $\text{ch}(B) \leqq b(\bar{B})$. Let $\bar{B}_1, \cdots, \bar{B}_k$ be interval graphs whose intersection is $\bar{B}$. The complements $B_1, \cdots, B_k$ of the $\bar{B}_i$'s are subgraphs of B and therefore are bipartite. Thus, the B_i's are chain subgraphs of B that cover its edges. □

COROLLARY 5. *It is* NP*-complete to determine if the boxicity of a graph is at most* 3.

Cozzens showed recently the NP-completeness of the boxicity problem for arbitrary k, i.e., that given graph G and number k it is NP-complete to tell if $b(G) \leqq k$ [C]. The boxicity 2 case remains open.

Cubicity. A *unit-interval graph* is the intersection graph of unit intervals (closed intervals of length 1) on the real line. The *cubicity* $c(G)$ of a graph G is the minimum number of unit-interval graphs whose intersection is G. Geometrically, the cubicity of G is the minimum number k such that G is the intersection graph of unit cubes with sides parallel to the coordinate axes in the k-dimensional space [R]. Clearly, $b(G) \leqq c(G)$.

LEMMA 6. *Let* $\bar{B}$ *be the complement of a bipartite graph* B. *Then,* $c(\bar{B}) = \text{ch}(B)$.

Proof. In view of Lemma 5 it suffices to show that the complement of a chain graph G is a unit interval graph. Let G be a chain graph that has the form of Fig. 1. We shall construct a unit-interval model for G. Associate with every node of R_i $(i = 1, \cdots, k)$ the (closed) unit interval $[i/k, 1 + i/k]$, and with every node of R_i' the interval $[1 + i/k, 2 + i/k]$. It is easy to see then that the intersection graph of these intervals is $\bar{G}$, the complement of G. □

COROLLARY 6. *It is* NP*-complete to determine if the cubicity of a graph is at most* 3. □

Threshold dimension. Let G be a graph with nodes $v_1, \cdots, v_n$. With every subset X of nodes we can associate its characteristic vector $\mathbf{x} = \langle x_1, \cdots, x_n \rangle$, where x_i is 1 or 0 depending on whether the node v_i is in X or not. The *threshold dimension* $\theta(G)$ of G is the minimum number of linear inequalities in the variables $x_1, \cdots x_n$ such that a set of nodes X is independent (i.e., does not induce any edge) if and only if its characteristic vector satisfies the inequalities [CH1]. A graph G with $\theta(G) \leqq 1$ is called a *threshold graph*. The threshold dimension of a graph G can be defined in an equivalent way as the minimum number of threshold subgraphs of G that cover its edges.

A threshold graph has the following structure. Its nodes can be partitioned into an independent set of nodes P and a clique Q so that the subgraph of G consisting of the edges of G between P and Q is a chain graph. Equivalently, G is a threshold graph if and only if it does not contain as an induced subgraph a pair of independent edges, a path of length 3, or a cycle of length 4 (see [CH1], [G] for more details).

LEMMA 7. *Let* B *be a bipartite graph with* P, Q *a bipartition of its nodes. Let* B' *be obtained from* B *by including all edges between nodes in* Q *(i.e., making* Q *a clique). Then* $\text{ch}(B) = \theta(B')$.

Proof. (1) $\text{ch}(B) \geqq \theta(B')$. Let $B_1, \cdots, B_k$ be chain subgraphs of B that cover its edges, and let $B_1', \cdots, B_k'$ respectively be obtained from them by turning Q into a clique. Then the B_i's are threshold subgraphs of B' that cover its edges.

(2) $\text{ch}(B) \leqq \theta(B')$. Let $B'_1, \cdots, B'_k$ be threshold subgraphs of B' that cover its edges. For each i, let B_i consist of the edges of B'_i between P and Q. Then the B_i's cover the edges of B. We claim that they are also chain graphs. For, suppose that B_i has a pair of independent edges $[x, y]$, $[u, v]$ with $x, u \in P$ and $y, v \in Q$. Then the subgraph of B'_i induced by these four nodes is either a path of length 3 (if it contains the edge $[y, v]$) or a pair of independent edges (if it does not contain $[y, v]$). In either case B'_i is not a threshold graph. □

COROLLARY 7. *It is* NP*-complete to determine if a given graph has threshold dimension at most* 3.

Chvatal and Hammer [CH2] had shown the NP-completeness of the threshold dimension problem in the case of arbitrary dimension. The case of dimension 2 remains open.

Acknowledgment. I wish to thank Martin Golumbic for pointing out the implication for the threshold dimension problem (Corollary 7) and for helpful discussions on the problems of § 4.

REFERENCES

[CH1] V. CHVÁTAL AND P. L. HAMMER, *Set-packing and threshold graphs*, Univ. Waterloo Res. report, CORR 73-21, Waterloo, Ontario, Canada 1973.

[CH2] ———, *Aggregation of inequalities in integer programming*, Ann. Discrete Math., 1 (1977), pp. 145–162.

[C] M. B. COZZENS, *Higher and multi-dimensional analogues of interval graphs*, Ph.D. thesis, Rutgers University, New Brunswick, NJ, 1981.

[DM] B. DUSHNIK AND E. W. MILLER, *Partially ordered sets*, Amer. J. Math., 63 (1941), pp. 600–610.

[F] P. C. FISHBURN, *Intransitive indifference with unequal indifference intervals*, J. Math. Psych., 7 (1970), pp. 144–149.

[GJ] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability, a Guide to the Theory of* NP*-Completeness*, W. H. Freeman, San Francisco, 1979.

[GJS] M. R. GAREY, D. S. JOHNSON AND L. STOCKMEYER, *Some simplified* NP*-complete graph problems*, Theoret. Comp. Sci. 1 (1976) pp. 237–267.

[G] M. C. GOLUMBIC, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[H] T. HIRAGUCHI, *On the dimension of partially ordered sets*, Sci. Rep. Konazowa Univ., 1 (1951), pp. 77–94.

[IP] T. IBARAKI AND U. N. PELED, *Sufficient conditions for graphs with threshold number* 2, Ann. Discrete Math., to appear.

[K] D. KELLY, *The* 3*-irreducible partially ordered sets*, Canad. J. Math., 29 (1977), pp. 367–383.

[Ki] R. KIMBLE, *Extremal problems in dimension theory for partially ordered sets*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, 1973.

[L] E. LAWLER, private communication

[O] O. ORE, *Theory of Graphs*, AMS Colloquium Publications 38, American Mathematical Society, Providence, RI, 1962.

[PLM] A. PNUELI, A. LEMPEL AND S. EVEN, *Transitive orientation of graphs and identification of permutation graphs*, Canad. J. Math., 23 (1971), pp. 160–175.

[R] F. S. ROBERTS, *On the boxicity and cubicity of a graph*, in Recent Progress in Combinatorics, W. T. Tutte, ed., Academic Press, New York (1969), pp. 301–310.

[TB] W. T. TROTTER AND K. P. BOGART, *On the complexity of posets*, Discrete Math., 16 (1976), pp. 71–82.

[TM] W. T. TROTTER JR AND J. I. MOORE JR., *Characterization problems for graphs, partially ordered sets, lattices and families of sets*, Discrete Math., 16 (1976), pp. 361–381.

[Y] M. YANNAKAKIS, *Computing the minimum fill-in is* NP*-complete*, this Journal, 2 (1981), pp. 77–79.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 3, September 1982

0196-5212/82/0303-0011 $01.00/0

MULTISETS OF APERIODIC CYCLES*

N. G. DE BRUIJN† AND D. A. KLARNER‡

Abstract. The basic result is that if A is a finite set then there are exactly $|A|^n$ multisets of aperiodic cycles over A with total length n. This is shown by a counting technique but also by establishing an explicit bijection from these multisets to words over A of length n.

1. Notation. Let $\mathbb{N}=\{0, 1, \cdots\}$ and $\mathbb{P}=\{1, 2, \cdots\}$ be the sets of nonnegative and positive integers respectively. Let A be a finite or countable set, and for each $n \in \mathbb{P}$ let A^n be the set of all n-tuples over A. An element $x \in A^n$ is called an *n-word* or sometimes just a *word*, and the *length* of x is defined to be $\lambda(x)=n$. Also, it is convenient to have an *empty word* Λ whose length is defined to be $\lambda(\Lambda)=0$. Let A^* be the set of all words over A, then elements $x, y \in A^*$ are concatenated to form a new word $xy \in A^*$ in the usual way. In particular, $\Lambda x = x\Lambda = x$ for all $x \in A^*$. Also, for each $k \in \mathbb{P}$, $x \in A^*$, x^k is defined to be the concatenation of k copies of x. (That is, $x^1 = x$, and $x^{k+1} = x \cdot x^k$ for all $k \in \mathbb{P}$). If $x, y \in A^*$ are such that $y = xu$ for some $u \in A^*$, we write $x \subseteq y$ and say x is an *initial word* of y.

Suppose $x \in A^n$, $n \in \mathbb{P}$, with $x = x_1 \cdots x_n$, $x_i \in A$ for $i = 1, \cdots, n$. Then the word $x_{d+1} \cdots x_n x_1 \cdots x_d$ is called the *d-shift* of x for $d = 1, \cdots, n-1$. If x is equal to a d-shift of x, then x is called *periodic*; otherwise, x is called *aperiodic*. The empty word is neither periodic, nor aperiodic. Suppose x is periodic and let $d \in \mathbb{P}$ be the smallest number such that x is equal to the d-shift of x. Then d is called the *period* of x and we define $\pi(x)$ to be the initial word of x with length d. If x is aperiodic we define $\pi(x) = x$. It is easy to check that $\pi(x)$ is an aperiodic word for all $x \in A^* \backslash \{\Lambda\}$. Also, if x is an n-word and $\pi(x)$ is a d-word where $n \in \mathbb{P}$, then d divides n. Furthermore, if $k = n/d$, then $x = (\pi(x))^k$. This motivates the definition of a *k-fold word*, namely, a word having the form y^k with y an aperiodic word. Let $P_k(A)$ be the set of k-fold words over A; in particular, $P_1(A)$ is the set of aperiodic words.

Now we define an equivalence relation on $A^* \backslash \{\Lambda\}$. Put $x \sim y$ just when $x = y$ or y is a d-shift of x for $1 \leqq d < \lambda(x)$. An equivalence class is called a *cycle* over A, and $\mathscr{C}(A)$ is defined to be the set of all cycles over A. If $x \in A^* \backslash \{\Lambda\}$, let $\langle x \rangle$ be the cycle which has x as an element. We will speak of a property common to all the words in an equivalence class as a property of the cycle, and notation will be used in a similar fashion even though this is a bit improper. For example, if $x \sim y$, then $\lambda(x) = \lambda(y)$. So we speak of the length of $\langle x \rangle$ and write $\lambda\langle x \rangle$ for it. Also, if $x \sim y$, then x and y have the same period, and $\pi(x) \sim \pi(y)$. So we speak of the period of $\langle x \rangle$, and write $\pi\langle x \rangle = \langle \pi(x) \rangle$. If x is a k-fold word, $\langle x \rangle$ is called a *k-fold cycle*. Let $\mathscr{A}_k(A)$ be the set of k-fold cycles over A, $k \in \mathbb{P}$. If x is aperiodic, we say $\langle x \rangle$ is aperiodic. Let $\mathscr{A}(A) = \mathscr{A}_1(A)$ be the set of aperiodic cycles over A. Thus, $\{\mathscr{A}_1(A), \mathscr{A}_2(A), \cdots\}$ is a partition of $\mathscr{C}(A)$ into disjoint sets.

A *finite multiset* on a set X is a mapping f from X into $\mathbb{N}$ such that the *size* of f, which is defined to be $\omega(f) = \sum_{x \in X} f(x)$, is finite. In this paper we shall just say *multiset* instead of "finite multiset". Let $\mathscr{M}(A)$ and $\mathscr{M}_k(A)$ be the set of all multisets on $\mathscr{C}(A)$ and $\mathscr{A}_k(A)$ respectively for all $k \in \mathbb{P}$. For f in $\mathscr{M}(A)$ or in $\mathscr{M}_k(A)$ we introduce

* Received by the editors July 14, 1981, and in revised from November 19, 1981.

† Eindhoven University of Technology, Department of Mathematics, 5600 MB Eindhoven, the Netherlands.

‡ Mathematical Sciences Department, State University of New York, Binghamton, New York, 13901.

a measure $\kappa(f)$ called the length total of f, that is different from the size $w(f)$. Instead of just counting the elements of the multiset, we give each element C a weight that equals $\lambda(C)$. So the *length total* of $f \in \mathcal{M}(A)$ is defined to be $\kappa(f) = \sum_{C \in \mathcal{C}(A)} f(C)\lambda(C)$, and $\kappa(f)$ for $f \in \mathcal{M}_k(A)$ is defined similarly as a sum over $\mathcal{A}_k(A)$. If $\kappa(f) = n$, we call f an *n-multiset* for all $n \in \mathbb{N}$.

2. Statement of results. The heart of our results is a bijection between the set of n-multisets over $\mathcal{A}(A)$ and the set of n-words over A. That is, every multiset of aperiodic cycles whose lengths total n corresponds to a word of length n, and conversely. We will use this bijection to prove a weighted version. Let w be a product weight on A^*. That is, w is a mapping of A^* into some sort of commutative algebra, and one of the most important properties of w is that $w(xy) = w(x)w(y)$ for all $x, y \in A^*$. This means $w(x_1 \cdots x_n) = w(x_1) \cdots w(x_n)$ for all $x_i \in A$, $i = 1, \cdots, n$. For our purposes, w must possess some other properties which guarantee that certain infinite sums and products are themselves weights. These extra assumptions will become evident later, but will not be explicitly stated. Since $x \sim y$ implies $w(x) = w(y)$, we speak of the weight of a cycle $C \in \mathcal{C}(A)$ and write $w(C) = w(x)$ for all $x \in C$.

Finally we define a weight function W on $\mathcal{M}(A)$ and $\mathcal{M}_k(A)$ in terms of w. For all $f \in \mathcal{M}(A)$ we put

$$W(f) = \prod_{C \in \mathcal{C}(A)} (w(C))^{f(C)}. \tag{1}$$

The weight of $f \in \mathcal{M}_k(A)$ is defined similarly as a product over $\mathcal{A}_k(A)$, $k \in \mathbb{P}$. Note that since $\{\mathcal{A}_1(A), \mathcal{A}_2(A), \cdots\}$ is a partition of $\mathcal{C}(A)$, a multiset $f \in \mathcal{M}(A)$ is completely determined by its restrictions to the blocks of this partition. Let f_k be the restriction of f to $\mathcal{A}_k(A)$ for all $k \in \mathbb{P}$. Then it is easy to see that

$$W(f) = \prod_{k=1}^{\infty} W(f_k). \tag{2}$$

and furthermore,

$$\sum_{f \in \mathcal{M}(A)} W(f) = \prod_{k=1}^{\infty} \sum_{f_k \in \mathcal{M}_k(A)} W(f_k). \tag{3}$$

Finally, we will prove that for all $k \in \mathbb{P}$

$$\sum_{f \in \mathcal{M}_k(A)} W(f) = \sum_{g \in \mathcal{M}_1(A)} (W(g))^k. \tag{4}$$

Here we are dealing with multisets of k-fold cycles. Recall that there is a natural bijection π between $\mathcal{A}_k(A)$ and $\mathcal{A}_1(A)$. Say $\langle x \rangle \in \mathcal{A}_k(A)$, then $\langle \pi(x) \rangle$ is the corresponding element in $\mathcal{A}_1(A)$. Going the other way, $\langle y \rangle \in \mathcal{A}_1(A)$ corresponds to $\langle y^k \rangle$ in $\mathcal{A}_k(A)$. Thus, $g \in \mathcal{M}_1(A)$ corresponds to $f \in \mathcal{M}_k(A)$ where $g\langle x \rangle = f\langle x^k \rangle$ for all $\langle x \rangle \in \mathcal{A}_1(A)$; also, $W(f) = (W(g))^k$ for all $k \in \mathbb{P}$.

We will prove in subsequent sections that

$$\sum_{f \in \mathcal{M}_1(A)} W(f) = \sum_{x \in A^*} w(x) = \sum_{n=0}^{\infty} \left(\sum_{a \in A} w(a) \right)^n. \tag{5}$$

However, an immediate consequence of (5) is that

$$\sum_{f \in \mathcal{M}_1(A)} (W(f))^k = \sum_{n=0}^{\infty} \left(\sum_{a \in A}^{\infty} (w(a))^k \right)^n. \tag{6}$$

This follows just by replacing w with the kth power of w in the definition of the weight of $f \in \mathcal{M}_1(A)$ so that W is replaced with the kth power of W. Combining (3), (4), and (6) gives

$$\sum_{f \in \mathcal{M}(A)} W(f) = \prod_{k=1}^{\infty} \sum_{n=0}^{\infty} \Big(\sum_{a \in A} (w(a))^k \Big)^n. \tag{7}$$

However, we still have to prove (5); this is done combinatorially in § 4 and algebraically in § 5.

3. Some motivation. The inspiration for this paper was a formula due to Read [8] which expresses $f(n)$, the number of endomorphism patterns on an n-set, in terms of $t(1), \cdots, t(n)$, where $t(k)$ is the number of isomorphism classes of rooted trees with n points. The formula is most elegantly expressed in terms of the two generating functions involved. Let $F(z) = 1 + f(1)z + f(2)z^2 + \cdots$, and let $T(z) = t(1)z + t(2)z^2 + \cdots$. Then Read's formula is

$$F(z) = \prod_{k=1}^{\infty} \frac{1}{1 - T(z^k)}. \tag{8}$$

Read's derivation of (8) was based on a formula due to Harary [6] who used Pólya's fundamental enumeration theorem to find an algorithm for computing the number of endomorphism patterns with n vertices.

An endomorphism on a set of n points can be described by means of a collection of cycles where each point of each cycle is the root of a tree. The idea that gave rise to the present paper is the discovery that the separate factors in (1) allow simple interpretations. The factor $(1 - T(z))^{-1}$ is related to the cycles on which the set of trees does not show any periodicity, and in general the factor $(1 - T(z^k))^{-1}$ is related to the cycles on which the set of trees has the exact period m/k (where m is the number of points on the cycle). A description of how endomorphisms are related to cycles with trees is given in § 6.

A quite simple observation is that the counting argument (presented in its simplest form in § 5) does not make use of the fact that the objects growing on the cycles are trees. We can replace the trees by the elements of any arbitrary finite or countable set A. This gives rise to the problem formulation presented in § 2.

4. The bijection. Our objective in this section is to construct a bijection Ω between the set of n-words over A and the set of n-multisets of aperiodic cycles over A represented in a certain normal form. Also, Ω preserves weights. That is, if x corresponds to multiset f, then $w(x) = W(f)$. Since

$$\sum_{x \in A^n} w(x) = \Big(\sum_{a \in A} w(a) \Big)^n,$$

we can sum this over all $n \in \mathbb{N}$ and use the bijection Ω to get (5).

The basis of Ω is an algorithm which factors an n-word into its corresponding n-multiset expressed in normal form. The algorithm and normal form both depend on an arbitrary linear order imposed on A. So we suppose the countable set A is ordered linearly by $\leqq$, and extend $\leqq$ to the lexicographical order on A^* in the usual way. That is, for all $x, y \in A^*$, $x \leqq y$ means $x \subseteq y$ or there exist $u, r, s \in A^*$ and $p, q \in A$ with $p < q$ such that $x = upr$, $y = uqs$.

We should not be trapped into thinking that for $u, v, x \in A^*$, the inequality $u < v$ always implies $ux < vx$. The implication is correct, however, if $\lambda(u) = \lambda(v)$. We shall use this repeatedly.

A word $x \in A^*$ is called *normal* if $\lambda(x) = 1$, or if $\lambda(x) > 1$ and x is less than all its d-shifts for $d = 1, \cdots, \lambda(x) - 1$. If x is aperiodic, then x and all its d-shifts $(d = 1, \cdots, \lambda(x) - 1)$ are distinct.

Let $N(A)$ be the set of normal words over A. Thus, every aperiodic cycle contains exactly one normal word, and every normal word x gives rise to an aperiodic cycle x. The sequence of normal words $(c_1, \cdots, c_k)$ is defined to be the *normal form* $\nu(f)$ of a multiset f over $\mathscr{A}(A)$ just when $c_1 \geqq \cdots \geqq c_k$, $\kappa(f) = \lambda(c_1) + \cdots + \lambda(c_k)$, and for each $c \in N(A)$ the number of i $(1 \leqq i \leqq k)$ with $c_i = c$ equals $f(c)$. Roughly speaking, normal words replace aperiodic cycles, and f is represented as a decreasing list of words after this replacement. For example, suppose $A = \{a, b, c, \cdots\}$ with $a < b < c < \cdots$, and define f to be 0 for all aperiodic cycles over A except that $f\langle a\rangle = 3$, $f\langle b\rangle = 1$, $f\langle ab\rangle = 2$, and $f\langle abacb\rangle = 1$. This means $\kappa(f) = 3 + 1 + 4 + 5 = 13$, and since b, $abacb$, ab, a are all normal and in decreasing order, $\nu(f) = (b, abacb, ab, ab, a, a, a)$. It will turn out that the word corresponding to f in this case is $\Omega^{-1}(f) = babacbababaaa$.

Let $x \in A^*$ with $x \neq \Lambda$, let $\sigma(x)$ be the longest normal initial word of x, and let $\tau(x)$ be the rest of x after $\sigma(x)$ has been deleted. Create a sequence $\Omega(x)$ of normal words as follows. Define $\Omega(x) = (x)$ for all normal words x, and define $\Omega(x) = (\sigma(x), \Omega(\tau(x)))$ for all other nonempty words. (Delete extra parentheses according to the rules $(u, (v)) = (u, v)$.)

If $\Omega(x) = (x_1, \cdots, x_n)$ then $x = x_1 \cdots x_n$. We shall show in Lemmas 3 and 4 that the converse is true if $x_1, \cdots, x_n$ are all normal and $x_1 \geqq \cdots \geqq x_n$.

LEMMA 1. *Suppose $x, y \in A^*$, y is normal, $x \neq \Lambda$ and $x < y$. Then $xy < yx$.*

Proof. By definition of $x < y$, there are two cases. In the first case, we have $y = xt$ with $t \in A^*$, $t \neq \Lambda$. Then $xt < tx$ because y is normal, so $xy = xxt < xtx = yx$. In the second case, we have $x = upr$, $y = uqs$ with $u, r, s \in A^*$, $p, q \in A$, and $p < q$. Then $xy = upry < uqsx = yx$ because $p < q$. This completes the proof. □

The next result is a generalization of this one.

LEMMA 2. *Suppose $x, y_1, \cdots, y_k \in A^*$, $y_1, \cdots, y_k$ normal, and $x \leqq y_1, \cdots, y_k$. Then $xy_1 \cdots y_k \leqq y_1 \cdots y_k x$.*

Proof. Using the previous lemma we have

$$xy_1 \cdots y_k \leqq y_1 xy_2 \cdots y_k \leqq y_1 y_2 x \cdots y_k \leqq \cdots \leqq y_1 y_2 \cdots y_k x. \tag{9}$$

That is, the ith inequality holds because $x \leqq y_i$ and because y_i normal implies $xy_i \leqq y_i x$ for $i = 1, \cdots, k$. This completes the proof. □

LEMMA 3. *Suppose $x_1, \cdots, x_h \in A^*$, $x_1, \cdots, x_h$ are normal, and $x_1 \geqq \cdots \geqq x_h$. Let $x = x_1 \cdots x_h$. Then $\Omega(x) = (x_1, \cdots, x_h)$. That is, every n-multiset over $\mathscr{A}(A)$ in normal form is the image under Ω of some element of A^n.*

Proof. It is enough to show that $\sigma(x) = x_1$, because then $\Omega(x) = (x_1, \cdots, x_n)$ follows by a simple induction. Since x_1 is normal, we certainly have $x_1 \subseteq \sigma(x)$. Suppose $\sigma(x) = x_1 \cdots x_k u$ where $u \subseteq x_{k+1}$ for some $k = 1, \cdots, h-1$, and $u \neq \Lambda$. Then $x_1, \cdots, x_k \geqq x_{k+1} \geqq x$, and $x_1, \cdots, x_k$ are normal by hypothesis, so $ux_1 \cdots x_k \leqq x_1 \cdots x_k u$ by Lemma 2. This means $x_1 \cdots x_k u$ is not normal. Thus, $\sigma(u) \subseteq x_1$, so $\sigma(u) = x_1$. This completes the proof. □

LEMMA 4. *Suppose $x \in A^*$, $x \neq \Lambda$, $\Omega(x) = (x_1, \cdots, x_k)$. Then $x_1 \geqq \cdots \geqq x_k$.*

Proof. It can be assumed without loss of generality that $k = 2$. If $k = 1$, there is nothing to prove. If the theorem is true for $k = 2$, then for $k \geqq 3$ one can use the fact that $\Omega(x) = (x_1, \cdots, x_k)$ implies $\Omega(x_1 x_{i+1}) = (x_i, x_{i+1})$ for $i = 1, \cdots, k-1$ to conclude that $x_i \geqq x_{i+1}$ for $i = 1, \cdots, k-1$; that is, $x_1 \geqq \cdots \geqq x_k$.

We will show that if x_1 and x_2 are normal words, and if $\Omega(x_1x_2) = (x_1, x_2)$, then $x_1 \geqq x_2$. This will be done by induction on the length of x_1x_2. Actually we shall prove for $n \geqq 2$ the following statement: for every linearly ordered set A, and for every pair x_1, x_2 of normal words over A with $\lambda(x_1x_2) = n$, $\Omega(x_1, x_2) = (x_1, x_2)$, we have $x_1 \geqq x_2$. If $n = 2$ this statement is true. Next we assume $n > 2$.

Let a_1, a_2 be the initial elements of x_1, x_2 respectively. Then because x_i is normal, every element of x_i is not less than a_i for $i = 1, 2$. Also, $a_1 \geqq a_2$, for if $a_1 < a_2$ we will show that x_1a_2 is normal, contradicting the assumption that $\sigma(x_1x_2) = x_1$. To show that x_1a_2 is normal if $a_1 < a_2$, we have to show that x_1a_2 is exceeded by all its shifts. First, $x_1a_2 < a_2x_1$ because $a_1 < a_2$. If $\lambda(x_1) = 1$ this shows that x_1a_2 is normal. Next, suppose $x_1 = uv$ with $u \neq \Lambda$, $v \neq \Lambda$. Then $uv < vu$ because x_1 is normal. Hence, $uva_2 < vua_2$. But the initial element of u is a_1, so $ua_2 < a_2u$. Hence, $uva_2 < vua_2 < va_2u$. This shows x_1a_2 is normal, a contradiction, so $a_1 \geqq a_2$.

If $a_1 > a_2$, we have $x_1 > x_2$ and we are done.

Finally we get to the hardest case: $a_1 = a_2 = a$. Then a is the smallest element in x_1x_2. Thus, we can put $x_1 = au_1 \cdots au_h$, $x_2 = av_1 \cdots av_j$ where the u's and v's are words either empty or with every element greater than a.

Let A_1 be the set of all $p \in A$ with $p > a$. Then words with every element greater than a are elements of A_1^*, and so is the empty word. The combinations ax with $x \in A_1^*$, will be called *syllables*. The $au_1, \cdots, au_h, av_1, \cdots, av_j$ mentioned in the previous paragraph are syllables.

Let us use the letter B for the set of all syllables. If $b_1, b_2 \in B$ we write $b_1 < b_2$ if and only if this inequality holds in A^* (elements of B are words over A). By lexicographic order, this inequality is extended to elements of B^* (the set of words over B).

There is a natural injection from words over B to words over A. For example, if $au_1, \cdots, au_k$ are syllables, then $(au_1)(au_2) \cdots (au_k)$ is a word over B, and it is mapped onto $au_1au_2 \cdots au_k$, which is a word over A. This injection is easily seen to preserve lexicographic order: $(au_1) \cdots (au_k) < (av_1) \cdots (av_j)$ in B^* if and only if $au_1 \cdots au_k < av_1 \cdots av_j$ in A^*. And it preserves normality: $(au_1) \cdots (au_k)$ is normal in B^* if and only if $au_1 \cdots au_k$ is normal in A^*.

Now let $x_1 = au_1 \cdots au_k$ and $x_2 = av_1 \cdots av_j$ be normal in A^*, with $\lambda(x_1x_2) = n$ and $\Omega(x_1x_2) = (x_1, x_2)$. Then in B^* the words $y_1 = (au_1) \cdots (au_k)$ and $y_2 = (au_1) \cdots (av_j)$ are normal, with $\Omega(y_1y_2) = (y_1, y_2)$. But $\lambda(y_1y_2)$ is less than n (the case that all syllables in x_1x_2 have length 1 is easily dismissed because $n > 2$). So by the induction hypothesis we have $y_1 \geqq y_2$. Since the injection preserves order, we conclude $x_1 \geqq x_2$. This completes the proof. □

The foregoing lemmas lead to the following.

THEOREM 1. *Every $x \in A^n$ corresponds to exactly one n-multiset f over $\mathcal{A}(A)$ such that $\Omega(x) = \nu(f)$. Furthermore $w(x) = W(f)$.*

This concludes our combinatorial proof of (5). To illustrate the bijection, consider all words over $A = \{a, b, c\}$ with $a < b < c$ which have two as, one b and one c. These words together with their Ω-factorizations are:

x	$\Omega(x)$	x	$\Omega(x)$
aabc	(*aabc*)	*baac*	(*b, aac*)
aacb	(*aacb*)	*baca*	(*b, ac, a*)
abac	(*abac*)	*bcaa*	(*bc, a, a*)
abca	(*abc, a*)	*caab*	(*c, aab*)
acab	(*ac, ab*)	*caba*	(*c, ab, a*)
acba	(*acb, a*)	*abaa*	(*c, b, a, a*)

We have listed the words in this illustration in lexicographical order. It might be noticed that if each $\Omega(x)$ is viewed as a word over $N(A)$, then the list of Ω-factorizations is also in lexicographical order. An explanation for this is given by the following results.

LEMMA 5. *Let $x, y \in A^*$ with $x \leqq y$. Then $\sigma(x) \leqq \sigma(y)$.*

Proof. If $x \leqq y$, there are two cases to consider. In the first case, we have $y = xt$ for some $t \in A^*$. Then $\sigma(x) \subseteq \sigma(xt) = \sigma(y)$, so $\sigma(x) \leqq \sigma(y)$. In the second case we have $x = upr$, $y = uqs$, $u, r, s \in A^*$, $p, q \in A$, and $p < q$. If $u = \Lambda$, then $p \subseteq \sigma(x)$, $q \subseteq \sigma(y)$, so $\sigma(x) < \sigma(y)$. If $u \neq \Lambda$, then either $\sigma(x) \subseteq u$, or $up \subseteq \sigma(x)$. If $\sigma(x) \subseteq u$, then $\sigma(x) = \sigma(u) \subseteq \sigma(y)$ since $u \subseteq y$, so $\sigma(x) \leqq \sigma(y)$. If $up \subseteq \sigma(x)$, we will show uq is normal, so $uq \subseteq \sigma(y)$, and we have $\sigma(x) < \sigma(y)$. To see that $up \subseteq \sigma(x)$ implies uq normal, suppose $\sigma(x) = upt$, $t \in A^*$. Let $u = u_1u_2$ with $u_1 \neq \Lambda$, then $u_1u_2pt < u_2ptu_1$ because upt is normal. We finally show that replacing pt by q preserves this inequality. We write $u_1u_2 = vrw$ where $v, w \in A^*$, $r \in A$ and $\lambda(v) = \lambda(u_2)$. Since $u_1u_2pt < u_2ptu$, we have $vrwpt < u_2ptu_1$. We consider the two cases $v < u_2$ and $v = u_2$ separately. If $v < u_2$ then $vrwq < u_2qu_1$ (because of $\lambda(v) = \lambda(u_2)$), so $u_1u_2q < u_2qu_1$. If $v = u_2$ we have $r \leqq p$, and therefore $r < q$, whence $vrwq < u_2qu_1$, so again $u_1u_2q < u_2qu_1$. This means that uq is normal, and our proof is complete. □

THEOREM 2. *Suppose $x < y$. Then $\Omega(x) < \Omega(y)$.*

Proof. This is proved by induction on $\lambda(x)$. The case $\lambda(x) = 1$ is trivial. Suppose the theorem is true for all $x \in A^*$ with $\lambda(x) < n$ for some $n \geqq 2$. Let $x, y \in A^*$ with $x < y$, $\lambda(x) = n$. We know from the previous lemma that $\sigma(x) \leqq \sigma(y)$. If $\sigma(x) = \sigma(y)$, then $\tau(x) < \tau(y)$, and $\lambda(\tau(x)) < \lambda(x)$, so $\Omega(\tau(x)) < \Omega(\tau(y))$ by the induction hypothesis. Hence, $\Omega(x) = (\sigma(x), \Omega(\tau(x))) < (\sigma(x), \Omega(\tau(y))) = \Omega(y)$. If $\sigma(x) < \sigma(y)$, then $\Omega(x) = (\sigma(x), \Omega(\tau(x))) < (\sigma(y), \Omega(\tau(y))) = \Omega(y)$. This completes the proof. □

As an application of Theorem 2, we mention that if v and w are normal words over A, and if $v < w$, then $v^n < w$ for all $n \in \mathbb{P}$. For if v and w are viewed as single letters, and $v < w$, then $v \cdots v$ is lexicographically less than w.

We state without proof another curious property. Let $\varepsilon(x)$ be the longest normal terminal word of x for all $x \in A^*$, and let $\delta(x)$ be the rest of x after $\varepsilon(x)$ has been deleted. Define $\Omega'(x)$, a factorization of x into normal words, as follows. First, $\Omega'(x) = (x)$ if x is normal. Otherwise if x is not empty, $\Omega'(x) = (\Omega'(\delta(x)), \varepsilon(x))$, and extra parentheses are deleted as in the definition of Ω. The surprise is that $\Omega(x) = \Omega'(x)$!

5. Algebraic proof of (5). Recall that $P_1(A)$ is the set of aperiodic words over A, and that $\mathscr{A}(A)$ is the set of aperiodic cycles over A. Every aperiodic cycle of length n can give rise to n distinct aperiodic words (which are obtained by breaking the cycle open at one of the n possible places).

Because of the definition of W and w in § 2 we have

$$\sum_{f \in \mathscr{M}_1(A)} W(f) = \sum_{C \in \mathscr{A}(A)} \prod_{k=0}^{\infty} (w(C))^k. \tag{10}$$

We shall use the identity

$$\sum_{k=0}^{\infty} z^k = \exp\left\{\sum_{k=1}^{\infty} \frac{z^k}{k}\right\}. \tag{11}$$

Applying this with $z = w(C)$, (10) becomes

$$\sum_{f \in \mathscr{M}_1(A)} W(f) = \prod_{C \in \mathscr{A}(A)} \exp\left\{\sum_{k=1}^{\infty} \frac{(w(C))^k}{k}\right\} = \exp\left\{\sum_{k=1}^{\infty} \frac{1}{k} \sum_{C \in \mathscr{A}(A)} (w(C))^k\right\}. \tag{12}$$

We shall prove the identity

$$\sum_{k=1}^{\infty} \frac{1}{k} \sum_{C \in \mathscr{A}(A)} (w(C))^k = \sum_{m=1}^{\infty} \frac{1}{m} \sum_{y \in A^m} w(y). \tag{13}$$

The left-hand side can be written as

$$\sum_{k=1}^{\infty} \sum_{n=1}^{\infty} \frac{1}{kn} \sum_{\substack{x \in P_1(A) \\ \lambda(x)=n}} (w(x))^k, \tag{14}$$

since each aperiodic cycle C with length n corresponds to exactly n elements of $P_1(A)$.

Taking terms with the same value of kn together, we transform (14) into

$$\sum_{m=1}^{\infty} \frac{1}{m} \sum_{n|m} \sum_{\substack{x \in P_1(A) \\ \lambda(x)=n}} (w(x))^{m/n}. \tag{15}$$

Every word y with length m can be written uniquely as $y = x^{m/n}$, where n is a divisor of m and x is aperiodic. Conversely, if n divides m and if $x \in P_1(A)$, $\lambda(x) = n$, then $x^{m/n} \in A^m$. Therefore we can write (15) as the right-hand side of (13), just noting that $(w(x))^{m/n} = w(x^{m/n})$. This proves (13).

Since

$$\sum_{y \in A^n} w(y) = \left(\sum_{a \in A} w(a) \right)^n \tag{16}$$

we find that application of the exponential function to the right-hand side of (13) leads to

$$\sum_{k=0}^{\infty} \left(\sum_{a \in A} w(a) \right)^k$$

(cf. (11)), whence (12) and (13) lead to (5).

As a generalization of (13) we mention, with an extra parameter s,

$$\sum_{k=1}^{\infty} \frac{1}{k^{s+1}} \sum_{c \in \mathscr{A}(A)} \frac{(w(c))^k}{(\lambda(c))^s} = \sum_{y \in A^* \setminus \{\Lambda\}} \frac{w(y)}{(\lambda(y))^{s+1}}. \tag{17}$$

The case $s = 0$ is (13), but the case $s = -1$ looks pretty as well.

6. Some examples. Let us return to the problem we described in § 3. Let D be an n-set, $n \in \mathbb{P}$, let $S(D)$ be the set (and group) of all permutations of D, and let D^D be the set of all mappings of D into D. Elements of D^D are called *endomorphisms* of D. The (directed) *graph* of $f \in D^D$ has vertex set D and edge set $\{(d, f(d)): d \in D\}$. Elements $f, g \in D^D$ are defined to be *equivalent* if the graph of f is isomorphic to the graph of g. This means there exists $\gamma \in S(D)$ such that $\{(\gamma d, \gamma f(d)): d \in D\} = \{(d, g(d): d \in D\} = \{(\gamma d, g\gamma(d): d \in D\}$, that is, $\gamma f = g\gamma$, which is the same as $\gamma f \gamma^{-1} = g$. An equivalence class in D^D is called an *endomorphism pattern* of D, and $f(n)$ is defined to be the number of these patterns for any n-set D. Next, $t(n)$ is defined to be the number of isomorphic classes of rooted trees with vertex set any n-set V, $n \in \mathbb{P}$. Such a class is called a *rooted tree pattern*. Diagrams representing rooted tree patterns having fewer than six vertices are shown in Fig. 1. Diagrams representing endomorphism patterns of an n-set for $n = 1, 2, 3$ are shown in Fig. 2. Also, Fig. 3 indicates how an endomorphism pattern might be viewed as a multiset of cycles of rooted tree patterns.

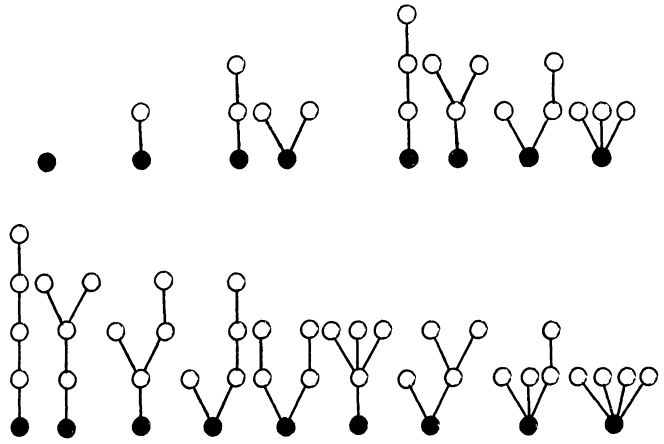


FIG. 1. *Rooted tree patterns.*

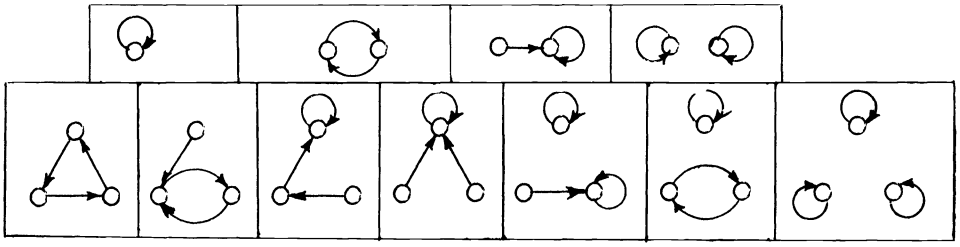


FIG. 2. *Endomorphism patterns.*

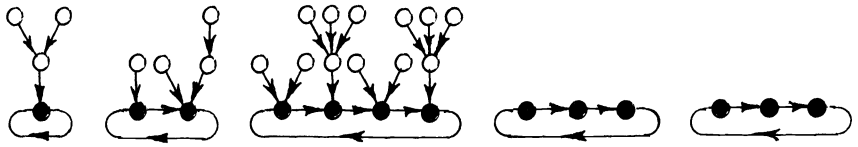


FIG. 3. *An endomorphism pattern viewed as a multiset of rooted tree cycles.*

Read's formula (8) tells us how to compute $f(n)$, the number of endomorphism patterns on an n-set, in terms of the rooted tree numbers. However, the bijection described in § 4 tells how to encode such a pattern as a sequence of words over the set of trees. The kth word in the sequence corresponds to the k-fold cycles of trees. For example, the endomorphism pattern in Fig. 3 is encoded as (6, 2, 7), (4, 14), (1, 1), $(\Lambda, \Lambda, \Lambda, \cdots)$. To decode a sequence of words $(x_1, x_2, \cdots)$, apply Ω to x_k for $k = 1, 2, \cdots$ to get $\Omega(x_k) = (x_{k1}, x_{k2}, \cdots)$; then x_{ki} is used to form a k-fold cycle of rooted trees $\langle x_{ki}^{k} \rangle$. We use as a convention that x_k is replaced by Λ if there are no k-fold cycles.

Our generalization of Read's formula allows us to enumerate other kinds of endomorphism patterns with equal ease. For example, suppose we are only interested in those $f \in D^D$ with $|f^{-1}(d)| \leqq h$; that is, every $d \in D$ is the image under f of at most h other elements of D, $h \in \mathbb{P}$. When $h = 1$, these endomorphisms are the permutations of D, and the patterns correspond to partitions of n if D is an n-set. For any $h \in \mathbb{P}$, endomorphisms f such that $|f^{-1}(d)| \leqq h$ give rise to patterns whose encoding as trees involves a special sort of rooted tree. Namely, the in-degree of the root vertex is at

most $h-1$, and the in-degrees of all other vertices are at most h. When $h=1$, there is only one tree like this, namely, the rooted tree with one vertex. In this case we would take $T(z)=z$ in (8) to get

$$\sum_{n=0}^{\infty} f(n)z^n = \prod_{k=1}^{\infty} \frac{1}{1-z^k}, \tag{18}$$

which is the generating function for the number of partitions of n, as expected. When $h=2$, the form of T is in part

$$T(z)=z+z^2+z^3+2z^4+3z^5+6z^6+11z^7+23z^8+46z^9+98z^{10}+\cdots.$$

We note that $T(z)=z(1+S(z))$, where $S(z)$ is the generating function for the rooted trees in which all vertices have degree $\leqq 2$. By Pólya's method (see [7]) we have

$$S(z)=z(1+S(z)+\tfrac{1}{2}S(z^2)+\tfrac{1}{2}(S(z))^2),$$

and therefore

$$T(z)=z+\tfrac{1}{2}(T(z^2)+(T(z))^2).$$

Using (8) with this new generating function T we get

$$\begin{aligned} F(z)=1&+z+3z^2+6z^3+15z^4+31z^5+75z^6+164z^7\\ &+388z^8+887z^9+2092z^{10}+4884z^{11}+11599z^{12}+27443z^{13}\\ &+65509z^{14}+156427z^{15}+375263z^{16}+\cdots. \end{aligned}$$

Thus, there are exactly 887 endomorphism patterns on a 9-set, involving $f\in D^D$ such that $|f^{-1}(d)|\leqq 2$ for all $d\in D$.

We close with some comments on the several papers which have dealt with the computation of $f(n)$. Fisher (1942) [4] seems to be first, and the same article with some corrections and additions appears in [5] (1950). In his 1950 reprinting of his earlier paper, Fisher adds a note indicating that he was unaware of Pólya's enumeration method. Nevertheless, Fisher's method produces results which run parallel to what one would get using Pólya's method. Davis (1953) [3] was aware of Pólya's method, but he elected to give an explicit formula for $f(n)$ using "Burnside's lemma" which is now properly renamed the Cauchy–Frobenius theorem. (See de Bruijn [1].) Harary (1959) [6] touched on the problem of computing $F(z)$ in his enumeration of patterns of functional digraphs, and he used Pólya's method. Read (1959) [8] obtained (8) by simplifying a formula given in Harary's paper. Finally, de Bruijn (1972) [2] investigated endomorphism patterns using the group action $\rho(\gamma)f=\gamma f\gamma^{-1}$, finding new proofs for older results.

Note added in proof. We are indebted to D. Foata for pointing out that Theorem 1 was known in the context of the theory of free Lie algebras. The oldest reference seems to be to A. I. Širšov, *Subalgebras of free Lie algebras*, Mat. Sbornik N.S. 33 (75) (1953) pp. 441–452. For related results see G. Viennot, *Algèbres de Lie libres et monoïdes libres*, Lecture Notes in Mathematics 691, Springer-Verlag, Berlin, Heidelberg, New York, 1978.

REFERENCES

[1] N. G. DE BRUIJN, *A note on the Cauchy–Frobenius lemma*, Nederl. Akad. Wetensch. Proc. Ser. A, 82 (= Indag. Math., 41) (1979), pp. 225–228.

[2] ———, *Enumeration of mapping patterns*, J. Combin. Theory, 12 (1972), pp. 14–20.
[3] R. L. DAVIS, *The number of structures of finite relations*, Proc. Amer. Math. Soc., 4 (1953), pp. 486–495.
[4] R. A. FISHER, *Some combinatorial theorems and enumerations connected with the number of diagonal types of a latin square*, Ann. Eugenics, XI, pt. IV, (1942), pp. 395–401.
[5] ———, *Contributions to Mathematical Statistics*, John Wiley, New York, 1950, pp. 41.394a–41.401.
[6] F. HARARY, *The number of functional digraphs*, Math. Ann., 138 (1959), pp. 203–210.
[7] G. PÓLYA, *Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen*, Acta Math., 68 (1937), pp. 145–254.
[8] R. C. READ, *A note on the number of functional digraphs*, Math. Ann., 143 (1961), pp. 109–110.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 3, September 1982

0196-5212/82/0303-0012 $01.00/0

SOME MAXIMAL SOLUTIONS OF THE GENERALIZED SUBADDITIVE INEQUALITY*

C. J. K. BATTY† AND D. G. ROGERS‡

Abstract. For real numbers $g(n, r)$, $0<r<n$, the generalized subadditive inequality

$$h(n) \leqq h(r)+h(n-r)+g(n, r), \qquad 0<r<n,$$

arises in a variety of problems of combinatorial interest. We study maximal solutions of this inequality, that is, sequences $\{f(n): n>0\}$ defined recursively by

$$f(n)=\min_{0<r<n}\{f(r)+f(n-r)+g(n, r)\}, \qquad n \geqq 2.$$

1. Introduction. Let $G=\{g(n, r): 0<r<n; n \geqq 2\}$ be an array of real numbers. A sequence $h=\{h(n): n \geqq 1\}$ is said to be a generalized subadditive sequence with respect to G if it satisfies the generalized subadditive inequality

$$h(n) \leqq h(r)+h(n-r)+g(n, r), \qquad 0<r<n. \tag{1}$$

If h is such a sequence, then so is $\{h(n)-nh(1): n \geqq 1\}$. So, without loss of generality, we standardize these sequences by taking $h(1)=0$. Also, replacing $g(n, r)$ by $\min(g(n, r), g(n, n-r))$, there is no loss of generality in assuming that $g(n, r)=g(n, n-r)$ and then restricting attention to $0<r \leqq [n/2]$ in (1), so we shall do this. (Here, as usual $[x]$ denotes the integer part of x.)

We are particularly interested in maximal solutions of (1), that is sequences $f=\{f(n): n \geqq 1\}$ defined by

$$f(1)=0, \qquad f(n)=\min_{0<r \leqq [n/2]} f(n, r), \qquad n \geqq 2, \tag{2a}$$

where

$$f(n, r)=f(r)+f(n-r)+g(n, r), \qquad 0<r \leqq \left[\frac{n}{2}\right]. \tag{2b}$$

Thus f is a generalized subadditive sequence with respect to G and is maximal among all such standardized sequences h in the sense that

$$h(n) \leqq f(n), \qquad n \geqq 1.$$

If f is the maximal solution of (1) with $f(1)=0$, then we write $f=T(G)$. So T transforms arrays G into sequences f according to (2). Furthermore for arrays G and H, and $a, b \geqq 0$,

$$T(aG+bH) \geqq aT(G)+bT(H), \tag{3}$$

where the addition and scalar multiplication of arrays and sequences is componentwise. We are also interested in the values of r for which the minimum in (2a) is achieved for given G and n, and we write

$$S(n; G)=\{r: f(n)=f(n, r)\}, \qquad n \geqq 2.$$

* Received by the editors May 8, 1980 and in revised form October 30, 1981.

† Department of Mathematics, University of Edinburgh, King's Buildings, Edinburgh EH9 3JZ, Scotland.

‡ 68 Liverpool Road, Watford, Hertfordshire, WD1 8DN, England.

Now if $S(n) = S(n; G) \cap S(n; H)$ is nonempty for $n \geqq 2$, and $a, b > 0$, then

$$S(n) = S(n; aG + bH), \qquad n \geqq 2 \tag{4}$$

and equality holds in (3). Thus T is linear on suitable cones of arrays.

Generalized superadditive sequences may be defined similarly. Indeed, if h is such a sequence with respect to the array G, then $-h = \{-h(n): n \geqq 1\}$ is a generalized subadditive sequence with respect to $-G = \{-g(n, r): 0 < r \leqq [n/2]; n \geqq 2\}$. Notice also that, if we replace min by max in (2a), then $-f$ is the maximal solution of (1) where $g(n, r)$ is replaced by $-g(n, r)$.

This terminology extends that of Hammersley and Grimmett [5], who, motivated by a problem in the growth of random objects [6], were chiefly concerned with the case where $g(n, r)$ is independent of r. In this case, Hammersley considers the inequality in the context of cooperative phenomena (see [4] and the references to his work given there) and establishes in [3] a generalization of the limit theorem for subadditive sequences (see also [2], [8]). Morris [9] also examines an example of this kind in some work on sorting.

However, the generalized subadditive inequality arises naturally in the wider sense of (1) in various combinatorial settings: for example, in the design of electrical circuits [10], [11], in the packing of multicolored graphs [12], [13] and in a problem in game theory [7]. Maximal solutions are then important in providing upper bounds in these problems. Moreover, Hammersley and Grimmett devote the latter part of their paper [5] to the study of one special case when $g(n, r)$ does depend on r as well as n, namely $g(n, r) = g(n) + c\delta_{r,n-r}$ where $c = 2 \log n$ and $\delta_{i,j}$ is Kronecker's δ. There is in addition, some resemblance between (2) and equations occurring in control theory (with feedback). With the possibility of further applications also in mind, it seems useful to allow this more general dependence and so to extend the terminology in this way.

We begin, in § 2, by recalling the known results when $g(n, r)$ is independent of r, and by showing that the general case may sometimes be reduced to this special one. In §§ 3 and 4, we consider cases (occurring in the applications mentioned above) in which $g(n, r)$ is independent of n, and linear in both n and r, in the latter case obtaining explicit expressions for the maximal solutions of (1). If h is a subadditive sequence, that is $g(n, r) = 0$ in (1), then it is a standard result [8, Chapt. 7] that the limit

$$l = \lim_{n \to \infty} \frac{h(n)}{n}$$

exists with $-\infty \leqq l \leqq \infty$. This result also holds under weaker conditions on $g(n, r)$ (see [2], [3]). A general question of interest is: how does the asymptotic behavior of a generalized subadditive sequence h satisfying (1) depend on the array G? We find that for several of our explicit maximal solutions f of (1), the limit

$$\lim_{n \to \infty} \frac{f(n)}{n \log n}$$

exists and is finite. We also show in § 5, that for $g(n, r) = n + kr$, this limit is robust in that it remains unchanged if (2a) holds only for $n > m$ for some $m \geqq 1$. This is of particular interest in some of the combinatorial applications where it may be possible to improve on the upper bounds given by (2a) for small values of n, simply by inspection.

2. The case $g(n, r)$ independent of r. The following theorem is essentially due to Hammersley and Grimmett [5] who used part (iii) in establishing Harding's conjec-

ture [6]. They confined attention to the case where, for $n \geqq 2$, $S(n; G)$ contains only one integer r, and they did not consider the properties of $f = T(G)$.

THEOREM 1. *Let $g(n, r) = g(n)$ be independent of r, and let $f = T(G)$.*

(i) *If g is decreasing, then 1 is in $S(n; G)$, $n \geqq 2$.*

(ii) *If g is increasing and convex, then $[n/2]$ is in $S(n; G)$, $n \geqq 2$, and f is also convex.*

(iii) *If g is increasing and concave, then $\rho(n)$ is in $S(n; G)$, $n \geqq 2$, where, for $p \geqq 0$,*

$$\rho(2^p + q) = \begin{cases} 2^{p-1}, & 0 \leqq q \leqq 2^{p-1}, \\ q, & 2^{p-1} \leqq q \leqq 2^p. \end{cases} \tag{5}$$

(iv) *If g is increasing and nonnegative, then f is increasing.*

Moreover, if the monotonicity (resp. convexity, concavity) in (i) (*resp.* (ii), (iii)) *is strict, then there is only one r in $S(n; G)$, $n \geqq 2$.*

Proofs of generalizations of parts (ii) and (iii) concerning k-partite divisions of n with $k \geqq 2$ are given in [1]. These differ from those in [5]; they turn more on the properties of f, and do not restrict $S(n; G)$, although the cases when $S(n; G)$ is restricted may be deduced from them. The results of [1] are substantially more general than those given in Theorem 1. The class of $g(n, r) = g(n)$ for which $[n/2]$ is in $S(n; G)$, $n \geqq 2$, is exactly described. Also if g is independent of r, then T in effect transforms sequences into sequences, and some consideration of this is given in [1].

Occasionally it is possible to reduce cases when $g(n, r)$ depends on r to those where $g(n, r)$ is independent of r. Suppose that

$$\hat{g}(n) = \min_{0 < r \leqq [n/2]} g(n, r) = g(n, s_n), \qquad n \geqq 2,$$

where s_n is in $S(n; \hat{G})$, $n \geqq 2$. Then it is easily seen that $T(G) = T(\hat{G})$ and s_n is in $S(n; G)$, $n \geqq 2$.

3. The case $g(n, r)$ independent of n. The following is an analogue of Theorem 1. An example of parts (ii) and (iii) occurs in [7].

THEOREM 2. *Let $g(n, r) = g(r)$ be independent of n in the range of definition: $n \geqq 2r > 0$.*

(i) *If g is increasing, then 1 is in $S(n; G)$, $n \geqq 2$.*

(ii) *If g is decreasing and concave, then $[n/2]$ is in $S(n; G)$, $n \geqq 2$, and 1 is in $S(2^p + 1; G)$, $p \geqq 0$. Also $T(G)$ is concave.*

(iii) *If g is decreasing and convex, then $\sigma(n)$ and $\rho(n)$ are in $S(n; G)$, $n \geqq 2$, where*

$$\sigma(2^p + q) = q, \qquad 0 < q \leqq 2^p, \quad p \geqq 0 \tag{6}$$

and $\rho(n)$ is given by (5).

Moreover, if the respective properties of g hold strictly, then the only r in $S(n; G)$, $n \geqq 2$, are those given above.

Proof. We shall write $f = T(G)$.

(i) It is straightforward to verify that in this case f is given by

$$f(n) = (n-1)g(1), \qquad n \geqq 1,$$

and that 1 is in $S(n; G)$ (compare [5, p. 274]).

(ii) It is convenient to write $f(0) = g(0) = 0$ so that we may then also write

$$f(1) = f(0) + f(1) + g(0) = f(1, 0).$$

With this convention, 0 is in $S(1; G)$. Also $[n/2]$ is in $S(n; G)$ for $n = 2$ and 3. Now

suppose that $[n/2]$ is in $S(n; G)$ for $1 \leqq n < m$ for some $m > 3$, and take $0 < r \leqq [m/2]$, so also $r \leqq m - r$. Then

$$f(m, r) = f(r) + f(m-r) + g(r) = f\left(\left[\frac{r}{2}\right]\right) + f\left(\left[\frac{r+1}{2}\right]\right) + g\left(\left[\frac{r}{2}\right]\right) + f\left(\left[\frac{m-2}{2}\right]\right) + f\left(\left[\frac{m-r+1}{2}\right]\right) + g\left(\left[\frac{m-r}{2}\right]\right) + g(r). \tag{7}$$

Now if at least one of r and $m - r$ is even, then

$$\left[\frac{r}{2}\right] + \left[\frac{m-r}{2}\right] = \left[\frac{m}{2}\right], \qquad \left[\frac{r+1}{2}\right] + \left[\frac{m-r+1}{2}\right] = \left[\frac{m+1}{2}\right],$$

so on rearranging (7) and using the inductive hypothesis, we have

$$\begin{aligned} f(m, r) &\geqq f\left(\left[\frac{m}{2}\right]\right) + f\left(\left[\frac{m+1}{2}\right]\right) + g\left(\left[\frac{m-r}{2}\right]\right) + g(r) - g\left(\left[\frac{r+1}{2}\right]\right) \\ &= f\left(m, \left[\frac{m}{2}\right]\right) + g\left(\left[\frac{m-r}{2}\right]\right) + g(r) - g\left(\left[\frac{r+1}{2}\right]\right) - g\left(\left[\frac{m}{2}\right]\right) \end{aligned} \tag{8a}$$

$$\geqq f\left(m, \left[\frac{m}{2}\right]\right), \tag{9a}$$

since g is concave,

$$\left[\frac{m-r}{2}\right] \leqq \left[\frac{m}{2}\right], \quad r \leqq \left[\frac{m}{2}\right], \quad \left[\frac{m-r}{2}\right] + r = \left[\frac{m+r}{2}\right] = \left[\frac{r+1}{2}\right] + \left[\frac{m}{2}\right].$$

If r and $m - r$ are both odd, then

$$\left[\frac{r}{2}\right] + \left[\frac{m-r+1}{2}\right] = \left[\frac{m}{2}\right], \qquad \left[\frac{r+1}{2}\right] + \left[\frac{m-r}{2}\right] = \left[\frac{m+1}{2}\right],$$

so again rearranging (7) and using the inductive hypothesis, we have

$$\begin{aligned} f(m, r) &\geqq f\left(\left[\frac{m}{2}\right]\right) + f\left(\left[\frac{m+1}{2}\right]\right) + g\left(\left[\frac{m-r}{2}\right]\right) + g(r) \\ &\qquad - g\left(\min\left(\left[\frac{r+1}{2}\right], \left[\frac{m-r}{2}\right]\right)\right) \\ &= f\left(m, \left[\frac{m}{2}\right]\right) + g\left(\left[\frac{m-r}{2}\right]\right) + g(r) \\ &\qquad - g\left(\min\left(\left[\frac{r+1}{2}\right], \left[\frac{m-r}{2}\right]\right)\right) - g\left(\left[\frac{m}{2}\right]\right) \end{aligned} \tag{8b}$$

$$\geqq f\left(m, \left[\frac{m}{2}\right]\right) \tag{9b}$$

since g is concave and decreasing,

$$\left[\frac{m-r}{2}\right] \leqq \left[\frac{m}{2}\right], \quad r \leqq \left[\frac{m}{2}\right], \quad \left[\frac{m-r}{2}\right] + r \leqq \min\left(\left[\frac{r+1}{2}\right], \left[\frac{m-r}{2}\right]\right) + \left[\frac{m}{2}\right].$$

Thus $[m/2]$ is in $S(m; G)$. It follows, by induction, that $[n/2]$ is in $S(n; G)$, $n \geqq 2$. Hence r is in $S(m; G)$ if and only if equality holds in both (8) and (9). It is now easy

to see by induction on p that 1 is in $S(2^p+1; G)$. Another inductive proof shows that f is concave (compare [10, p. 167]).

If now g is strictly concave, equality holds in (9) only if $r=[m/2]$ or $[(m-r)/2]=[m/2]$. But in the latter case, $r=1$ and m is odd; furthermore in this case, equality holds in (8a) only if 1 is in $S([(m+1)/2]; G)$. A further inductive argument shows that 1 is in $S(n; G)$ only if $n=2^p+1$ for some $p\geqq 0$. Now it follows that $S(n; G)$ is as stated.

(iii) The proof will again be by induction. It is readily verified that $\sigma(n)$ is in $S(n; G)$ for $n=2, 3$. Suppose that $\sigma(n)$ is in $S(n; G)$ for $2\leqq n<m$ for some $m>3$, and take $0<r\leqq[m/2]$. Let $p\geqq 1$ be the integer such that $2^p<m\leqq 2^{p+1}$, so that $\sigma(m)=m-2^p$. There are three cases to be considered.

First, suppose that $0<r<2^p<m-r<m$. Then $\sigma(m-r)=m-r-2^p$, so

$$f(m, r)=f(r)+f(m-r-2^p)+f(2^p)+g(m-r-2^p)+g(r)$$

$$\begin{aligned}(10a)\qquad &\geqq f(m-2^p)+f(2^p)-g(\min(r, m-r-2^p))+g(m-r-2^p)+g(r)\\ &=f(m, \sigma(m))-g(\min(r, m-r-2^p))+g(m-r-2^p)+g(r)-g(m-2^p)\end{aligned}$$

$$(11a)\qquad \geqq f(m, \sigma(m)),$$

since g is decreasing.

Second, suppose that $0<r\leqq 2^{p-1}<m-r\leqq 2^p<m$. Then $\sigma(m-r)=m-r-2^{p-1}$, $\sigma(m-2^{p-1})=m-2^p$ and $\sigma(2^p)=2^{p-1}$, so

$$f(m, r)=f(r)+f(m-r-2^{p-1})+f(2^{p-1})+g(m-r-2^{p-1})+g(r)$$

$$\begin{aligned}(10b)\qquad &\geqq f(m-2^{p-1})+f(2^{p-1})-g(\min(r, m-r-2^{p-1}))+g(m-r-2^{p-1})+g(r)\\ &=f(m-2^p)+2f(2^{p-1})+g(m-2^p)-g(\min(r, m-r-2^{p-1}))\\ &\quad+g(m-r-2^{p-1})+g(r)\\ &=f(m-2^p)+f(2^p)+g(m-2^p)-g(\min(r, m-r-2^{p-1}))\\ &\quad+g(m-r-2^{p-1})+g(r)-g(2^{p-1})\\ &=f(m, \sigma(m))-g(\min(r, m-r-2^{p-1}))+g(m-r-2^{p-1})+g(r)-g(2^{p-1})\end{aligned}$$

$$(11b)\qquad \geqq f(m, \sigma(m)),$$

since g is decreasing.

Third, suppose that $2^{p-1}<r\leqq m-r\leqq 2^p<m$. Then $\sigma(r)=r-2^{p-1}$, $\sigma(m-r)=m-r-2^{p-1}$ and $\sigma(2^{p-1})=2^p$, so

$$f(m, r)=f(r-2^{p-1})+f(m-r-2^{p-1})+2f(2^{p-1})+g(r-2^{p-1})+g(m-r-2^{p-1})+g(r)$$

$$\begin{aligned}(10c)\qquad &\geqq f(m-2^p)+f(2^p)-g(2^{p-1})+g(m-r-2^{p-1})+g(r)\\ &=f(m, \sigma(m))-g(2^{p-1})+g(m-r-2^{p-1})+g(r)-g(m-2^p)\end{aligned}$$

$$(11c)\qquad \geqq f(m, \sigma(m)),$$

since g is convex. It follows from these three cases that $\sigma(m)$ is in $S(m; G)$. Furthermore r is in $S(m; G)$ if and only if equality holds in both (10) and (11).

If $2^p<m\leqq 3.2^{p-1}$ and $r=\rho(m)=2^{p-1}$, the second case applies, and equality holds in (10b) and (11b). If $3.2^{p-1}<m\leqq 2^{p+1}$ and $r=\rho(m)=\sigma(m)$, the third case applies and equality certainly holds in (10c) and (11c). This shows that $\rho(m)$ is in $S(m; G)$.

If g is strictly convex, strict inequality holds in (11c) unless $r=m-2^p=\sigma(m)$. Also g is strictly decreasing, so strict inequality always holds in (11a); also strict

inequality holds in (11b) unless $r = m - 2^p = \sigma(m)$ or $r = 2^{p-1} = \rho(m)$. Thus $S(m; G) = \{\rho(m), \sigma(m)\}$, $m \geqq 2$.

4. The case $g(n, r)$ linear in n and r. In this section, we obtain some explicit expressions for the maximal solutions of (1) in some cases where $g(n, r)$ is linear in n and r for $0 < r \leqq [n/2]$; $n \geqq 2$. Thus let $h(n)$, $n \geqq 1$, be the sum of the coefficients in the binary expansion of n and let $H(n)$ and $J(n)$, $n \geqq 2$, be given by

$$H(n) = \sum_{i=1}^{n-1} h(i), \qquad n \geqq 1, \tag{12a}$$

$$J(2^p + q) = p2^p + q(p+2), \qquad 0 \leqq q \leqq 2^p, \quad p \geqq 0. \tag{12b}$$

Then, as an application of earlier results, we have:

THEOREM 3. *Let $g(n, r) = an + br$, $0 < r \leqq [n/2]$; and let $f = T(G)$.*

(i) *$a < 0 < b$, then $S(n; G) = \{1\}$, $n \geqq 2$, and f is given by*

$$f(n) = \frac{a}{2}(n(n+1) - 2) + b(n-1), \qquad n \geqq 1. \tag{13}$$

(ii) *If $b < 0 < a$, then $S(n; G) = \{r: \rho(n) \leqq r \leqq [n/2]\}$, $n \geqq 2$, where $\rho(n)$ is given by* (5). *Further f is given by*

$$f(n) = aJ(n) + bH(n), \qquad n \geqq 1. \tag{14}$$

Proof. (i) Let $g_1(n, r) = -n$; $g_2(n, r) = r$, $0 < r \leqq [n/2]$; $G_j = g_j(n, r)$: $0 < r \leqq [n/2]\}$; $f_j = T(G_j)$, $j = 1, 2$. So $G = (-a)G_1 + bG_2$. By Theorems 1 and 2, $S(n; G_1) = S(n; G_2) = \{1\}$, $n \geqq 2$. By (4), $S(n; G) = \{1\}$. It is easy to verify by induction that $f_1(n) = -\frac{1}{2}(n(n+1) - 2)$; $f_2(n) = n - 1$, $n \geqq 1$. Also equality holds in (3) (with a replaced by $-a$), so $f(n)$ is given by (13).

(ii) Let $g_1(n, r) = n$; $g_2(n, r) = -r$, $0 < r \leqq [n/2]$, so $G = aG_1 + (-b)G_2$. By Theorems 1 and 2, $\rho(n)$ and $[n/2]$ belong to both $S(n; G_1)$ and $S(n; G_2)$, $n \geqq 2$. It is easy to verify by induction that $f_1(n) = J(n)$; $f_2(n) = -H(n)$, $n \geqq 1$, and (compare Theorem 4 below)

$$S(n; G_1) = \{r: \rho(n) \leqq r \leqq [n/2]\}, \qquad S(n; G_2) = \{r: \sigma(n) \leqq r \leqq [n/2]\}, \qquad n \geqq 2$$

(where $\sigma(n)$ is given by (6)). By (4),

$$S(n; G) = S(n; G_1) \cap S(n; G_2) = \left\{r: \rho(n) \leqq r \leqq \left[\frac{n}{2}\right]\right\}, \qquad n \geqq 2.$$

Also equality holds in (3) (with b replaced by $-b$), so $f(n)$ is given by (14).

In the limiting case when $b = 0$, the behavior of $f = T(G)$ in Theorem 3 is given by Theorem 1; if $a = 0$, it is given by Theorem 2 (see also [1], [7], [9]). However if a and b have the same sign, the behavior is more complicated. The case when $a > 0$ and $b = ka$ for some integer $k \geqq 0$ arises in several combinatorial applications [9], [12], [13], and fortunately we can obtain an explicit solution for f as in Theorem 4 below. Similar behavior is observed if k is nonintegral, and it would be interesting to generalize Theorem 4.

Let k be a fixed nonnegative integer, and let $a_k(p) = a_p$, $p \geqq 1$, be given by:

$$a_p = 1, \qquad 1 \leqq p \leqq k+1, \tag{15a}$$

$$a_{p+1} = a_p + a_{p-k}, \qquad p > k. \tag{15b}$$

Note that, taking $k=0$, $a_0(p)=2^{p-1}$, and Theorem 4 reduces to a known result (compare (12), (14) and [1], [9]); taking $k=1$ gives an interesting occurrence of the Fibonacci numbers as the sequence $a_1(p)$, $p \geqq 1$.

THEOREM 4. *Let $g(n,r)=n+kr$, $0<r\leqq[n/2]$, where k is a fixed nonnegative integer, and let $f=T(G)$. Then for $p>k+1$, $0\leqq q\leqq a_{p-k}$,*

$$S(a_p+q;G)=\left\{r:\max(a_{p-k-1},q)\leqq r\leqq\min\left(a_{p-k},a_{p-k-1}+q,\left[\frac{a_p+q}{2}\right]\right)\right\} \tag{16}$$

and

$$f(a_p+q)=pa_p-a_{p+k}+(p+1)q, \tag{17}$$

where a_p, $p\geqq 1$, is given by (15).

Proof. Once again the proof is by induction. It is easy to verify that (16) and (17) hold for $p=k+2$ (and $q=0$ or 1). Suppose (16) and (17) hold for $k+1<p<m$, where $m>k+2$. Then in particular, using (15b),

$$S(a_m;G)=S(a_{m-1}+a_{m-k-1};G)=\{a_{m-k-1}\},$$

$$f(a_m)=f(a_{m-1}+a_{m-k-1};G)=(m-1)a_{m-1}-a_{m+k-1}+ma_{m-k-1}=ma_m-a_{m+k}.$$

Thus (16) and (17) are satisfied for $p=m$ and $q=0$. Suppose they hold for $p=m$ and $0\leqq q<j$ where $j>0$. For $0<r<[(a_m+j)/2]$, let

$$h(r)=f(a_m+j,r+1)-f(a_m+j,r).$$

Then by the inductive hypotheses and (17) (note that (17) is also valid for $p=0$; $q=0$ or 1),

$$h(r)=p(r)-p(a_m+j-r-1)+k,$$

where $p(r)$ is the integer greater than k such that $a_{p(r)}\leqq r<a_{p(r)+1}$, etc. Since $p(r)$ increases with r, so does $h(r)$, and $f(a_m+j,r)$ attains its minimum at any value of r where $h(r)=0$.

If $0<j<a_{m-k-1}$ (so that $m>2k+2$), then $p(a_{m-k-1})=m-k-1$ and $p(a_m+j-a_{m-k-1}-1)=m-1$, so $h(a_{m-k-1})=0$. Furthermore, $h(r)=0$ if and only if $p(r)=m-k-1$ and $p(a_m+j-r-1)=m-1$, i.e.,

$$\max(a_{m-k-1},j)\leqq r<\min(a_{m-k},a_{m-k-1}+j).$$

Thus (16) holds for $p=m$; $q=j$. Furthermore by the inductive hypotheses and (17),

$$\begin{aligned}f(a_m+j)&=f(a_m+j,a_{m-k-1})\\&=(m-k-1)a_{m-k-1}-a_{m-1}+(m-1)a_{m-1}-a_{m+k-1}+mj+a_m+j+ka_{m-k-1}\\&=m(a_{m-k-1}+a_{m-1})+(a_m-a_{m-k-1}-a_{m-1})-(a_{m-1}+a_{m+k-1})+(m+1)j\\&=ma_m-a_{m+k}+(m+1)j,\end{aligned}$$

using (15b). Thus (17) holds for $p=m$, $q=j$.

Very similar calculations show (subject to the inductive hypotheses) that if $a_{m-k-1}\leqq j<a_{m-k}$ (so that $m>2k+1$), then (16) and (17) hold for $p=m$; $q=j$.

If $m>2k+1$ and $j=a_{m-k}$, we have

$$h(a_{m-k})=m-k-(m-1)+k=1, \qquad h(a_{m-k}-1)=m-k-1-m+k=-1.$$

Hence (17) holds for $p = m$; $q = j$. Furthermore,

$$\begin{aligned} f(a_m + a_{m-k}) &= f(a_m + a_{m-k}, a_{m-k}) \\ &= (m-k)a_{m-k} - a_m + ma_m - a_{m+k} + a_m + a_{m-k} + ka_{m-k} \\ &= ma_m - a_{m+k} + (m+1)a_{m-k}, \end{aligned}$$

using (17) and (15b). Thus (17) also holds for $p = m$; $q = j$.

Finally if $k+2 < m \leqq 2k+1$ and $j = 1 = a_{m-k}$, we have

$$h(1) = k + 1 - (m-1) + k > 0.$$

Hence $S(a_m + 1; G) = \{1\} = \{a_{m-k}\}$, so (16) is valid for $p = m$; $q = j$. Furthermore,

$$\begin{aligned} f(a_m + 1) = f(a_m + 1, 1) &= ma_m - a_{m+k} + a_m + 1 + k \\ &= ma_m - a_{m+k} + m + 1, \end{aligned}$$

since $a_m = m - k$. Thus (17) is also valid for $p = m$, $q = j$. It now follows by induction that (16) and (17) hold for all stated values of p and q.

5. Robustness of the asymptotic behavior. It is easy to see that if $H(n)$ and $J(n)$ are given by (12), then

$$\lim_{n\to\infty} \frac{H(n)}{n \log_2 n} = \frac{1}{2}, \qquad \lim_{n\to\infty} \frac{J(n)}{n \log_2 n} = 1.$$

Thus, if f is given by (14),

$$\lim_{n\to\infty} \frac{f(n)}{n \log_2 n} = a + \frac{b}{2}.$$

If a_p, $p \geqq 1$, is given by (15) and γ is the unique root of $x^{k+1} - x^k - 1 = 0$ larger than 1, it is routine to verify that $a_{p+1}/a_p \to \gamma$, so that $(\log a_p)/p \to \log \gamma$ as $p \to \infty$. Hence if f is given by (17),

$$\lim_{n\to\infty} \frac{f(n)}{n \log n} = \frac{1}{\log \gamma}. \tag{18}$$

We shall now see that (18) remains valid if f is a maximal generalized subadditive function for $g(n, r) = n + kr$, subject to different initial conditions from those in (2a).

THEOREM 5. *Let $k \geqq 0$ and $m \geqq 1$ be integers, and $h = \{h(n): n \geqq 1\}$ be a sequence satisfying*

$$h(n) = \min\left\{h(r) + h(n-r) + n + kr: 0 < r \leqq \left[\frac{n}{2}\right]\right\}, \qquad n > m.$$

Then

$$\lim_{n\to\infty} \frac{h(n)}{n \log n} = \frac{1}{\log \gamma}$$

where γ is the largest real root of $x^{k+1} - x^k - 1 = 0$.

Proof. Let f be the sequence given by (17), and put $f(0) = 0$, so that by Theorem 4

$$f(x) + f(y) + x + y + kx \geqq f(x + y) \tag{19}$$

for any nonnegative integers x and y. Extend f to a function defined on all nonnegative

real numbers by taking f to be linear on the intervals $[n, n+1]$, $n \geqq 0$. Then it follows by linearity, first in x and then in y (noting that f is convex on $[0, \infty)$), that (19) is valid for all $x, y \geqq 0$. Furthermore (18) and some simple estimates give

$$\lim_{x\to\infty} \frac{f(x)}{x \log x} = \frac{1}{\log \gamma}.$$

A simple inductive argument shows that

$$mf\left(\frac{n}{m}\right) + \alpha n \leqq h(n) \leqq f(n) + \beta n, \qquad n \geqq 1,$$

where $\alpha = \min\{h(n)/n : 1 \leqq n \leqq m\}$; $\beta = \max\{h(n)/n : 1 \leqq n \leqq m\}$. The theorem now follows easily.

M. J. Pelling has observed that Theorem 5 remains valid even if k is not an integer.

Example. Bilinear arrays G arise in various graph-packing problems (see [12], [13]). For integers $m \geqq 3$, $n \geqq 1$, let $\mathcal{G}_m(n)$ be the class of edge-colored graphs Γ with n colors such that

(i) every complete graph on m vertices contained in Γ has exactly two colors,

(ii) for any pair of colors, there is a complete graph on m vertices contained in Γ colored with these two colors.

Let $p_m^*(n)$ be the minimal number of edges in a graph in $\mathcal{G}_m(n)$.

For $0 < r \leqq [n/2]$, let H_{mnr} be a graph on $r(m-2)+2$ vertices, constructed as follows. There are two distinguished vertices v_1 and v_2, joined by $n-r$ edges, one of each of the last $n-r$ colors; the remaining vertices are divided into r disjoint sets V_i, $1 \leqq i \leqq r$, each containing $m-2$ vertices; each vertex in V_i is joined to v_1, v_2 and each other vertex of V_i by an edge of color i. Taking Γ to be the disjoint union of H_{mnr}, a graph in $\mathcal{G}_m(r)$ colored with the first r colors, and a graph in $\mathcal{G}_m(n-r)$ colored with the last $(n-r)$ colors, Γ belongs to the class $\mathcal{G}_m(n)$. Hence

$$\begin{aligned} p_m^*(n) &\leqq p_m^*(r) + p_m^*(n-r) + n - r + (\tfrac{1}{2}m(m-1) - 1)r \\ &= p_m^*(r) + p_m^*(n-r) + n + kr, \end{aligned}$$

where $k = \frac{1}{2}m(m-1) - 2$. Since $p_m^*(1) = 0$, an upper bound for $p_m^*(n)$ is given by (17). It should be noted however that this bound is not precise. For example, (17) gives $p_3^*(3) \leqq 7$, whereas it may be seen directly that $p_3^*(3) = 6$. Theorem 5 gives

$$\limsup_{n\to\infty} \frac{p_3^*(n)}{n \log n} \leqq \frac{1}{\log \gamma},$$

where $\gamma = \frac{1}{2}(1+\sqrt{5})$, but no lower bound has been obtained for the lim inf.

REFERENCES

[1] C. J. K. Batty, M. J. Pelling and D. G. Rogers, *Some recurrence relations of recursive minimization*, this Journal, 3 (1982), pp. 13–29.

[2] N. G. de Bruijn and P. Erdös, *Some linear and some quadratic recursion formulas*, I, II, Indag. Math., 13 (1951), 374–382; 14 (1952), 145–163.

[3] J. M. Hammersley, *Generalization of the fundamental theorem on subadditive functions*, Proc. Cambridge Phil. Soc., 58 (1962), 235–238.

[4] ———, *Postulates for subadditive processes*, Ann. Probability, 2 (1974), 652–680.

[5] J. M. Hammersley and G. R. Grimmett, *Maximal solutions of the generalized subadditive inequality*, in Stochastic Geometry, E. F. Harding and D. G. Kendall, eds., John Wiley, London, 1974, pp. 270–284.

[6] E. F. HARDING, *The probabilities of the shapes of randomly bifurcating trees*, in Stochastic Geometry, E. F. Harding and D. G. Kendall, eds., John Wiley, London, 1974, pp. 259–269.
[7] S. HART, *A note on the edges of the n-cube*, Discrete Math., 14 (1976), 157–163.
[8] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, revised edition, AMS Colloquium Publications 31, American Mathematical Society, Providence, RI, 1957.
[9] R. MORRIS, *Some theorems on sorting*, SIAM J. Appl. Math., 17 (1969), pp. 1–6.
[10] M. J. PELLING AND D. G. ROGERS, *A problem in the design of electrical circuits*, in Combinatorial Mathematics V: Proceedings of the Fifth Australian Conference, Lecture Notes in Mathematics 622, Springer-Verlag, Berlin, 1977, pp. 153–169.
[11] ———, *Further results on a problem in the design of electrical circuits*, in Combinatorial Mathematics: Proceedings of the International Conference, Canberra, 1977, Lecture Notes in Mathematics 686, Springer-Verlag, Berlin, 1978, pp. 240–247.
[12] D. G. ROGERS, *Packing multicoloured graphs*, to appear.
[13] ———, *A packing problem for star graphs*, to appear.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 3, September 1982

0196-5212/82/0303-0013 $01.00/0

PARALLEL ALGORITHMS FOR NETWORK ROUTING PROBLEMS AND RECURRENCES*

JOHN A. WISNIEWSKI† AND AHMED H. SAMEH‡

Abstract. In this paper, we consider the parallel solution of recurrences, and linear systems in the regular algebra of Carré. These problems are equivalent to solving the shortest path problem in graph theory, and they also arise in the analysis of Fortran programs. Our methods for solving linear systems in the regular algebra are analogues of well-known methods for solving systems of linear algebraic equations. A parallel version of Dijkstra's method, which has no linear algebraic analogue, is presented. Considerations for choosing an algorithm when the problem is large and sparse are also discussed.

Key words. Parallel algorithms, shortest path problem, recurrence, networks, regular algebra, graphs

1. Introduction. A basic problem in applications of graph theory is that of finding shortest paths in a weighted graph. This problem arises in finding the shortest or least cost path in transportation problems, in solving minimum cost flow problems, in finding the critical (i.e., longest) path in scheduling problems, and in finding adversary routes through nuclear fuel-cycle facilities [13], [23].

While there are several versions of the shortest path problem [8], we will deal with the problem of finding the shortest paths in definite graphs G from one node to all others, where the path length used is the sum of the weights of the arcs in the path. A digraph $G(V, A)$ is a set V of nodes and a set A of arcs which are ordered pairs of nodes. Assume that the nodes are numbered from 1 to n so that $V = \{1, 2, \cdots, n\}$. Each arc $(i, j) \in A$ has an associated length, $\delta_{i,j}$. By defining $\delta_{i,i} = \infty$, $1 \leqq i \leqq n$ and $\delta_{i,j} = \infty$ for $(i, j) \notin A$, we have the n by n distance matrix $D = [\delta_{i,j}]$. If a graph G contains no cycles the sum of whose arc weights are less than or equal to zero, then we say that G is definite.

Although extensive work has been done on sequential algorithms for solving the shortest path problem, little work has been done [4], [6], [19] on parallel algorithms for this problem. Levitt and Kautz [19] discuss an iterative algorithm due to Hu [12] which has been tailored to a cellular array machine. Chen and Feng [4] discuss a version of Ford's algorithm [9], [10, pp. 130–133] for an associative processor machine. Dekel and Sahni [6] present a method for cube connected and perfect shuffle computers. Here, however, we treat the shortest path problem in a more general way, using the regular algebra of Carré [2]. Furthermore, we will assume that:

(i) any number of processors may be used at any time, but we will give bounds on this number;
(ii) each processor may perform either a comparison or any of the four arithmetic operations in one time step; and
(iii) there are no memory or data alignment time penalties.

If p processors are being used, then we denote the computation time as T_p unit steps. Thus T_1 is the time required by a serial machine. We define the speedup of a computation using p processors over the serial computation time as $Sp = T_1/T_p \leqq 1$.

Carré [2], has shown that the shortest path problem, as well as many other problems, can be posed as linear systems of the form $x = Ax + b$ in a regular algebra. He gives the examples of the scheduling algebra of Cruon and Hervé [5], the two

* Received by the editors August 8, 1980. This work was supported in part by the National Science Foundation under grant US NSF MCS75-21758, and by the U.S. Department of Energy.

† Division 2113, Sandia Laboratories, Albuquerque, New Mexico 87185.
‡ Department of Computer Science, University of Illinois, Urbana, Illinois 61801.

elements from boolean algebra, and a stochastic communication problem (Kalaba [16], Moisil [20]). Triangular linear systems, or recurrences, in a regular algebra also arise in the analysis of Fortran programs [17]. Consequently, it is important to solve efficiently such recurrences on a parallel computer.

In addition to considering the solution of linear systems in a regular algebra, we will present a parallel version of Dijkstra's algorithm [7] which is applicable only to graphs with positive arc weights. Furthermore, considerations for choosing an algorithm when the graph is sparse, i.e., has a large number of infinite entries in the distance matrix, are discussed.

1.1. Notation. We follow the convention that capital letters denote matrices, lower case letters denote vectors, and lower case greek letters denote scalars. Except in the cases where we state time and processor bounds, the symbols + and × are taken to represent the *generalized addition* and *generalized multiplication* operations of the regular algebra. In the statement of the time and processor bounds, the symbols + and × will take on their normal meaning. We will frequently omit the symbol × when it is clear that a generalized product is to be taken. We will also use the $\sum$ and Π notations for the generalized sum or product of a set of items. Unless otherwise stated, $\log n \equiv \lceil \log_2 n \rceil$ for any positive number n.

1.2. The regular algebra. For convenience, we restate the algebra of Carré. We start by defining a *semiring* $(S, +, \times)$ which satisfies the following properties:

(i) commutativity $\alpha+\beta=\beta+\alpha$; $\alpha\times\beta=\beta\times\alpha$;

(ii) associativity $\alpha+(\beta+\gamma)=(\alpha+\beta)+\gamma$; $\alpha\times(\beta\times\gamma)=(\alpha\times\beta)\times\gamma$;

(iii) distributivity $\alpha\times(\beta+\gamma)=(\alpha\times\beta)+(\alpha+\gamma)$;

(iv) idempotency $\alpha+\alpha=\alpha$

for α, β, γ, in S.

The set S has a unit element ε such that

$$\alpha\times\varepsilon=\alpha,$$

and a null element θ satisfying

$$\alpha+\theta=\alpha, \qquad \alpha\times\theta=\theta,$$

for all $\alpha \in S$. Furthermore, we have the law of multiplicative cancellation

(v) if $\alpha \neq \theta$ and $\alpha\times\beta=\alpha\times\gamma$ then $\beta=\gamma$.

We define for the semiring $(S, +, \times)$ the order relation $\leqq$

$$\alpha \leqq \beta \text{ if and only if } \alpha+\beta=\alpha.$$

1.3. Extensions to matrix operations. We now extend the definitions of the regular algebra to matrices all of whose elements belong to the set S. The definitions of matrix addition, matrix multiplication and matrix transposition are analogous to those in linear algebra. Note that matrix addition is idempotent,

$$A+A=A.$$

We define a square m by m unit matrix $I=[\varepsilon_{i,j}]$ with $\varepsilon_{i,j}=\varepsilon$ if $i=j$ and $\varepsilon_{i,j}=\theta$ if $i\neq j$. Note that $IA=AI=I$ for any m by m matrix A. The null matrix N, all of whose elements are θ, satisfies

$$A+N=A, \qquad A\times N=N.$$

The partial ordering for matrices is easily extended:

$$A\leqq B \text{ if and only if } \alpha_{i,j}\leqq\beta_{i,j} \quad \text{for all } i,j$$

In particular, it follows that

$$A\leqq B \quad \text{if and only if } A+B=A.$$

The shortest path problem can now be stated as a linear system in the regular algebra. Let $G(V,A)$ be a graph with distance matrix D. Let the vector b represent an initial labeling of the nodes V (for the vector b, β_i is the initial label on node i). The shortest path problem is equivalent to solving the linear system $x=Dx+b$ in the regular algebra [2]. For the single source problem the vector b is given by $\beta_s=\varepsilon$ where s is the source node, and $\beta_i=\theta$ otherwise.

2. Solution of recurrences. In this section we study the solution of systems of the form

$$x=Lx+f, \tag{1}$$

where the matrix L is a strictly lower triangular matrix. These systems arise in the analysis of Fortran programs, and are also used in methods for solving general systems

$$x=Ax+b.$$

2.1. The column sweep algorithm. The most fundamental of all algorithms for solving recurrences in the regular algebra is the column sweep algorithm. The solution of $x=Lx+f$, where

$$L=\begin{bmatrix} \theta & \theta & \cdots & \theta \\ \lambda_{2,1} & \theta & \cdots & \theta \\ \lambda_{3,1} & \lambda_{3,2} & \theta & \theta \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{n,1} & \lambda_{n,2} & \lambda_{n,n-1} & \theta \end{bmatrix}, \quad x=\begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \vdots \\ \xi_n \end{bmatrix}, \quad f=\begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \vdots \\ \phi_n \end{bmatrix} \tag{2}$$

is given by

$$\xi_1=\phi_1, \qquad \xi_j=\sum_{i=1}^{j-1}\lambda_{j,i}\xi_i+\phi_j, \qquad i=2,3,\cdots,n. \tag{3}$$

Rewriting the system (3) as column operations we get the

COLUMN SWEEP ALGORITHM.

1. Set $x\leftarrow f$.
2. For $i=1,2,\cdots,n-1$
   $x\leftarrow x+\xi_i l_i$ (in parallel)

In the algorithm, l_i represents the ith column of the matrix L. A sequential algorithm for solving recurrence (1) requires n^2-n operations, i.e., $T_1=n^2-n$. On the other hand, the column sweep algorithm can be performed in $2(n-1)$ time steps using at most $n-1$ processors. This gives a speedup of $n/2+O(1)$ over the sequential algorithm with a corresponding efficiency of roughly $\frac{1}{2}$.

2.2. The product form of the solution. In this section, the solution x of (1) is formed as a sum of products. Each product involves a power of the matrix L multiplied by the vector f.

LEMMA 1. *Let L be strictly lower triangular, i.e., L has the form given in equation (2). Also, let*

$$L^i = \begin{bmatrix} N & N \\ R^T & N \end{bmatrix},$$

where R is upper triangular of order $n-i$, i.e., L^i is lower triangular with i null diagonals in its lower half.

Proof. From the definition of matrix multiplication and the facts that $\theta+\theta=\theta$, $\theta\times\alpha=\theta$ for all $\alpha\in S$, and that L is null on the main diagonal, each multiplication of L^{i-1} by L introduces another diagonal of null elements into the product. □

COROLLARY 1. $L^n = N$.

LEMMA 2. *The product L^iL^j, $i \leqq j$, $i+j<n$ can be formed in* $\log(n-i-j)+1$ *steps using at most $p(n-i-j)$ processors, where*

$$p(k) = \sum_{i=1}^{k} \sum_{j=1}^{i} j = \frac{k(k+1)(k+2)}{6}.$$

Proof. This result also follows from the definition of matrix multiplication. Since L^i has i null diagonals and L^j has j null diagonals, the longest dot product formed is of length $n-i-j$. This establishes the time bound. The processor bound is established by counting the number of multiplications needed to compute the product. This is the stated processor count since all of the multiplications can be formed simultaneously. □

The product form of the solution is found by substituting for x in the right-hand side of (1), its equivalent, $Lx+f$. Thus,

$$x = L^2x + Lf + f.$$

After $n-1$ such substitutions, we obtain

$$x = L^nx + L^{n-1}f + \cdots + Lf + f.$$

From Corollary 1, $L^n = N$ so

$$x = L^{n-1}f + L^{n-2}f + \cdots Lf + f. \tag{4}$$

Furthermore, (4) can be written in the form

$$x = (I+L^{n/2})(I+L^{n/4})\cdots(I+L^2)(I+L)f. \tag{5}$$

Heller's algorithm [11] for solving triangular systems of linear algebraic equations, can now be applied.

ALGORITHM 1 (product form of solution for recurrences).

1. Set $x_0 \leftarrow f$.
2. For $i = 0, \cdots, \nu = \log(n/4)$
   $x_{i+1} \leftarrow (I+L^{2^i})x_i$,
   $L^{2^{i+1}} \leftarrow L^{2^i}L^{2^i}$.
3. $x \leftarrow (I+L^{n/2})x_{\nu+1}$.

THEOREM 1. *Algorithm 1 can be performed in parallel in* $\log^2 n + 3\log n + 1$ *steps using at most $\frac{1}{6}n^3$ processors.*

Proof. Heller has already shown that Algorithm 1 can be performed in $\log^2 n + 3\log n + 1$ steps. The formation of x_{i+1} and x requires only $O(n^2)$ processors whereas the formation of $L^{2^{i+1}}$ requires $O(n^3)$ processors. In particular, the largest number of processors are required in forming L^2. From Lemma 2, this number is

$$p(n-2) = \tfrac{1}{6}(n-2)(n-1)n < \tfrac{1}{6}n^3$$

processors. □

2.3. Product of elementary matrices solution. The method presented in this section is an analogue of Sameh and Brent's Algorithm I [22]. It is superior to Algorithm 1 above since it takes approximately half the time while using significantly fewer processors to solve the recurrence (1).

DEFINITION. M_j is an *elementary matrix*, if it is of the form,

$$M_j = \begin{bmatrix} \theta & & & & & & & \\ & \theta & & & & & N & \\ & & \ddots & & & & & \\ & & & \ddots & \theta & & & \\ N & & & & \mu_{j+1,j} & \theta & & \\ & & & & \mu_{j+2,j} & & \ddots & \\ & & & & \vdots & N & & \\ & & & & \mu_{n,j} & & & \theta \end{bmatrix},$$

M_j has all null entries except in column j below the main diagonal.

LEMMA 3. *Let M_i and M_j be elementary matrices, then $M_iM_j = N$ if $i \leqq j$. Furthermore if $i > j$ then $P = M_iM_j$ has all null elements except in column j below the ith row.*

Proof. M_j has its first j rows all null and M_i has all null elements except in column i below the main diagonal. Thus if $i \leqq j$ all inner products of rows of M_i with columns of M_j must be null, i.e., $M_iM_j = N$. If $i > j$ then the only nonnull inner products of $P = M_iM_j$ occur when the inner products of rows $i+1, i+2, \cdots, n$ of M_i are taken with column j of M_j. □

LEMMA 4. *Let $M_1, M_2, \cdots, M_j, M_{j+1}$ be elementary matrices. Then the product*

$$\begin{aligned}(I+M_{j+1})(I+M_j)\cdots(I+M_1)(I+M_1)(I+M_2)\cdots(I+M_{j+1}) \\ = (I+M_{j+1})(I+M_j)\cdots(I+M_1).\end{aligned} \tag{6}$$

The proof is by induction on j.

Basis step. For $j = 1$, we have

$$(I+M_1)(I+M_1) = I + M_1 + M_1 + M_1^2.$$

From Lemma 3 and idempotency

$$I + M_1 + M_1 + M_1^2 = I + M_1 + M_1 = I + M_1,$$

establishing the basis.

Induction step. Assume that the assertion holds for index j, and let

$$\bar{M}_j = (I+M_j)(I+M_{j-1})\cdots(I+M_2)(I+M_1).$$

From the inductive hypothesis we have

$$\begin{aligned}&(I+M_{j+1})(I+M_j)\cdots(I+M_1)(I+M_1)(I+M_2)\cdots(I+M_{j+1}) \\ &\quad = (I+M_{j+1})\bar{M}_j(I+M_{j+1}).\end{aligned}$$

Furthermore,

$$\bar{M}_j(I+M_{j+1}) = \bar{M}_j + \bar{M}_j M_{j+1}. \tag{7}$$

By repeated applications of Lemma 3 we obtain

$$\bar{M}_j M_{j+1} = M_{j+1},$$

hence (7) becomes

$$\bar{M}_j(I+M_{j+1}) = \bar{M}_j + M_{j+1}.$$

Consequently,

$$\begin{aligned}(I+M_{j+1})\bar{M}_j(I+M_{j+1}) &= (I+M_{j+1})(\bar{M}_j+M_{j+1})\\ &= \bar{M}_j + M_{j+1}\bar{M}_j + M_{j+1} + M_{j+1}^2\\ &= \bar{M}_j + M_{j+1}\bar{M}_j + M_{j+1}\\ &= \bar{M}_j + M_{j+1}(\bar{M}_j + I).\end{aligned}$$

Since I is an element of the expanded sum of $\bar{M}_j$, then due to idempotency, we have

$$(I+M_{j+1})\bar{M}_j(I+M_{j+1}) = \bar{M}_j + M_{j+1}\bar{M}_j = (I+M_{j+1})\bar{M}_j,$$

establishing the lemma. □

In a similar fashion, it can be shown that

$$\begin{aligned}(I+M_1)(I+M_2)\cdots(I+M_{j+1})(I+M_{j+1})(I+M_j)\cdots(I+M_1)\\ = (I+M_{j+1})(I+M_j)\cdots(I+M_1).\end{aligned} \tag{8}$$

Let

$$(I+L)^* \equiv (I+M_{n-1})(I+M_{n-2})\cdots(I+M_2)(I+M_1). \tag{9}$$

Since

$$(I+L) = (I+M_1)(I+M_2)\cdots(I+M_{n-2})(I+M_{n-1}),$$

from Lemma 4 and (8) we see that

$$(I+L)^*(I+L) = (I+L)^* = (I+L)(I+L)^*. \tag{10}$$

We now introduce a lemma of fundamental importance to what follows.

LEMMA 5. *The vector x solves the system $x = Ax+f$ if and only if x solves the system $x = x+Ax+f$.*

Proof. If $x = Ax+f$ then adding x to both sides and using idempotency gives

$$x+x = x+Ax+f, \qquad x = x+Ax+f.$$

If $x = x+Ax+f$ then

$$x+Ax+f \leqq Ax+f, \tag{11}$$

i.e., either $x = x$ or $x = Ax+f$. Since $x = x$ is redundant, $x = Ax+f$. □

The following result was originally given by Carré [2], but we state and prove it in terms of elementary matrices.

THEOREM 2. *Let $x = (I+L)^*f$, then x satisfies $x = Lx+f$.*

Proof. From Lemma 5, $x = Lx+f$ is equivalent to $x = (I+L)x+f$, or

$$x = (I+L)(I+L)^*f+f = (I+L)^*f+f.$$

From (9) f is an element of the expanded sum of $(I+L)^*f$, thus $x=(I+L)^*f$ satisfies (1). □

From Theorem 2, we can write the solution x of (1) as

$$x=(I+L)^*f=(I+M_{n-1})(I+M_{n-2})\cdots(I+M_2)(I+M_1)f,$$

which motivates the following algorithm.

ALGORITHM 2 (*solution of recurrences in product form*).
1. Set $(I+M_i)^{(0)}=I+M_i$, $i=1,\cdots,n-1$; $f^{(0)}=f$.
2. For $j=0,1,\cdots,\nu=\log(n/4)$
   Form $(I+M_i)^{(j+1)}=(I+M_{2i+1})^{(j)}(I+M_{2i})^{(j)}$
   in parallel for $i=1,2,\cdots,n/2^{j+1}-1$.
   Form $f^{(j+1)}=(I+M_1)^{(j)}f^{(j)}$.
3. Set $x=(I+M_1)^{(\nu)}f^{(\nu)}$.

Algorithm 2 is a direct analogue of Algorithm I of Sameh and Brent [22] for triangular linear algebraic equations. Hence, their results ((22, Thm. 2.1]) apply. The solution x to the recurrence (1) can be found in

$$\tau=\tfrac{1}{2}\log^2 n+\tfrac{3}{2}\log n+3$$

steps using no more than

$$\pi=\left(\frac{15}{1024}\right)n^3-O(n^2)\leqq\frac{n^3}{68}+O(n^2)$$

processors.

2.4. Limited parallelism. In this section we present two methods which are analogues of those of Hyafil and Kung [15] for solving the recurrence (1). These methods are used when the number of processors p is fixed.

The first method utilizes the algorithm decomposition applied to the column sweep algorithm:

$$x^{(1)}=f,\qquad x^{(i+1)}=(I+M_i)x^{(i)},\qquad i=1,2,\cdots,n-1.$$

It is easy to see that

$$(I+M_i)x^{(i)}=\begin{bmatrix}\xi_1^{(i)}\\ \vdots\\ \xi_i^{(i)}\\ \xi_{i+1}^{(i)}+\lambda_{i+1,i}\xi_i^{(i)}\\ \vdots\\ \xi_n^{(i)}+\lambda_{n,i}\xi_i^{(i)}\end{bmatrix}.$$

Thus, given p processors, the product $(I+M_i)x^{(i)}$ can be formed in $2\lceil(n-i)/p\rceil$ steps. Since $T_p(n)=\sum_{i=1}^{n}2\lceil(n-i)/p\rceil$ we have from Hyafil and Kung [15] that

$$T_p(n)\leqq\frac{n^2}{p}+\left(2-\frac{3}{p}\right)n+\frac{2}{p}-2.$$

This method is most practical when $p<n$.

The second method uses the problem decomposition principle. Let

$$L=\begin{bmatrix} L_{1,1} & & & \\ L_{2,1} & L_{2,2} & & N \\ \vdots & \vdots & \ddots & \\ L_{m,1} & L_{m,2} & \cdots & L_{m,m} \end{bmatrix}, \quad x=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}, \quad f=\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix},$$

then (1) can be written as,

$$\begin{aligned} x_1 &= L_{1,1}x_1 + f_1, \\ x_2 &= L_{2,1}x_1 + L_{2,2}x_2 + f_2, \\ &\vdots \\ x_m &= \Big(\sum_{j=1}^{m} L_{m,j}x_j\Big) + f_m. \end{aligned}$$

This leads to the following algorithm.

ALGORITHM 3. *Recurrence solver with limited number of processors.*
For $i = 1, \cdots, m$
  Set $f^{(i)} = \sum_{j=1}^{i-1} L_{i,j}x_j$.
  Form $x_i = (I + L_{i,i})^* f^{(i)}$ using Algorithm 2.

Hyafil and Kung have shown that for $p = \lceil n^r \rceil$, $1 < r < 3$, and taking

$$m = \left\lceil \frac{68n}{\lceil n^r \rceil^3} \right\rceil,$$

then

$$T_p(n) \leqq \begin{cases} O(n^{2-r} \log n) & \text{for } 1 < r < \frac{3}{2}, \\ O(n^{1-r/3} \log^2 n) & \text{for } \frac{3}{2} < r < 3. \end{cases}$$

3. Banded recurrences. In this section, we take the matrix L of (1) to be banded, i.e., of the form

$$L=\begin{bmatrix} L_1 & & & & & \\ R_1 & L_2 & & & & N \\ & R_2 & L_3 & & & \\ & & \ddots & \ddots & & \\ N & & & \ddots & \ddots & \\ & & & & \ddots & \ddots \\ & & & & R_{k-1} & L_k \end{bmatrix}, \tag{12}$$

where L_i is a strictly lower triangular matrix of order m, $i = 1, 2, \cdots, k$, and R_j is an upper triangular matrix of order m, $j = 1, 2, \cdots, k-1$.

3.1. Unlimited parallelism. We can show that the time and processors required for solving (1) are the same as given by Sameh and Brent [22]. The proof is exactly as that of their Theorem 3.1. Before this can be proved, however, we need to establish the following.

LEMMA 6. *Let L be a strictly lower triangular $2n$ by $2n$ matrix given by*

$$L=\begin{bmatrix} L_1 & N \\ R & L_2 \end{bmatrix}, \tag{13}$$

where R is n by n and upper triangular. Let f be a $2n$-vector correspondingly partitioned as

$$f^T = (f_1^T, f_2^T).$$

Thus, the solution of the linear recurrence

$$x = (I_{2n} + L)x + f \tag{14}$$

is given by

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} I_n & N \\ G & I_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \tag{15}$$

where

$$G = (I_n + L_2)^* R,$$

and

$$y_i = (I_n + L_i)^* f_i, \qquad i = 1, 2.$$

Proof. From the structure of L and Theorem 2, we see that

$$x_1 = (I_n + L_1)^* f_1 = y_1.$$

Also, from (13) and (14) we have

$$x_2 = Rx_1 + (I_n + L_2)x_2 + f_2.$$

From Theorem 2, we have

$$x_2 = (I_n + L_2)^*[R_1 x_1 + f_2] = Gy_1 + y_2,$$

proving the lemma. □

Now we state the main result.

THEOREM 3. *The linear recurrence*

$$x = (I + L)x + f,$$

where L is an n by n strictly lower triangular matrix (12), *i.e., $n = km$, is obtained in* $(2 + \log m)\log n - \frac{1}{2}(\log m)(1 + \log m) + 3$ *time steps requiring less than* $m(m+1)n/2$ *processors.*

Proof. Let $f^T = (f_1^T, f_2^T, \cdots, f_k^T)$. The algorithm of Lemma 6 can be generalized to yield a scheme which is exactly similar to that of Sameh and Brent [22, Thm. 3.1], as follows.

ALGORITHM 4. *Banded recurrences.*

1. Set $y_1^{(0)} = (I + L_1)^* f_1$, and

   $$[G_{i-1}^{(0)}, y_i^{(0)}] = (I + L_i)^*[R_{i-1}, f_i], \qquad i = 2, 3, \cdots, n/m.$$

2. For $j = 1, 2, \cdots, \nu = \log(n/m) - 1$,
   Set $r = 2^j m$.

   $$\text{Set } G_i^{(j+1)} = \begin{bmatrix} N & G_{2i}^{(j)} & \\ N & G_{2i+1}^{(j)} & G_{2i}^{(j)} \end{bmatrix}, \qquad i = 1, 2, \cdots, \frac{n}{2r} - 1, \text{ and}$$

   $$y_i^{(j+1)} = \begin{bmatrix} y_{2i-1}^{(j)} \\ G_{2i-1}^{(j)} y_{2i-1}^{(j)} + y_{2i}^{(j)} \end{bmatrix}, \qquad i = 1, 2, \cdots, \frac{n}{2r}.$$

Upon termination, $y_1^{(\nu+1)}$ contains the solution vector x.

3.2. Limited parallelism. In this section we will assume that $m < p \ll n$, where p is the number of processors available. If $p = m$ we can use the column sweep algorithm to solve the recurrence in $2(n-1)$ steps. We will develop a faster algorithm for $p > m$ along the same lines as the algorithm of Kuck and Sameh [18, Thm. 2.3.13], for solving banded triangular systems of linear algebraic equations. Our results are summarized by the following theorem.

THEOREM 4. *Given p processors, $m < p \ll n$, a linear recurrence with the matrix L of the form* (12) *can be solved in time*

$$T_p = 2(m-1) + \tau\left\lceil \frac{n-m}{p(p+m-1)} \right\rceil, \tag{16}$$

where

$$\tau = \max \begin{cases} (2m^2+3m)p - \dfrac{m}{2}(2m^2+3m+5) \\ 2m(m+1)p - 2m. \end{cases} \tag{17}$$

This yields a speedup and efficiency proportional to (p/m) *and* $(1/m)$, *respectively.*

Proof. To motivate the approach, consider a recurrence of the form

$$z = \hat{L}\hat{z} + g \quad \text{of order } s = q + p(p-1)$$

equations with $q = mp$,

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{bmatrix} = \begin{bmatrix} R_0 & L_1 & & & \\ & R_1 & L_2 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & R_{p-1} & L_p \end{bmatrix} \begin{bmatrix} z_0 \\ z_1 \\ z_2 \\ \vdots \\ z_p \end{bmatrix} + \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_p \end{bmatrix}. \tag{18}$$

Here, L_1 is of order q, $L_i (i > 1)$ is of order p, and $R_i (0 \leqq i \leqq p-1)$ contains nonnull elements only on its top m super diagonals, i.e.,

$$R_i = \begin{bmatrix} N & \tilde{R} \\ N & N \end{bmatrix}, \tag{19}$$

in which $\tilde{R}$ is upper triangular of order m. The system (18) can be expressed as

$$\begin{aligned} z_1 &= R_0 z_0 + L_1 z_1 + g_1, \\ z_2 &= R_1 z_1 + L_2 z_2 + g_2, \\ &\vdots \\ z_p &= R_{p-1} z_{p-1} + L_p z_p + g_p. \end{aligned}$$

From Theorem 2, we get that the z_is can be written as

$$\begin{aligned} &z_j = (I + L_j)^*[R_{j-1} z_{j-1} + g_j], \qquad j = 1, \cdots, p, \quad \text{or} \\ &z_j = ((I + L_j)^* R_{j-1}) z_{j-1} + (I + L_j)^* g_j. \end{aligned} \tag{20}$$

Therefore, we define

$$h_1 = (I + L_1)^*(R_0 z_0 + g_1), \tag{21}$$

$$[G_{i-1}, h_i] = (I + L_i)^*[R_{i-1}, g_i], \qquad i = 2, 3, \cdots, p. \tag{22}$$

Using a single processor, (21) can be evaluated in $\tau_1 = 2m^2p$ steps. Using a single processor to evaluate each of the $p-1$ systems, (22) can be evaluated in

$$\tau_2 = \left[(2m^2+m)p - \frac{m}{2}(2m^2+3m-1)\right]$$

steps.

Kuck and Sameh [18] have shown that the choice of $q = mp$ keeps the difference in time for evaluating (21) and (22) as small as practically possible. Thus the evaluation of (21) and (22) can be done in at most $\tau_3 = \max\{\tau_1, \tau_2\}$ steps.

We now see from (20), (21) and (22) that z is given by

$$z_1 = h_1, \qquad z_i = h_i + G_{i-1}z_{i-1}, \qquad i = 2, 3, \cdots, p.$$

Since only the last m columns of G_i are nonnull, by using all of the p processors, each z_i can be computed in $2m$ time steps. Thus, $z_2, z_3, \cdots, z_p$ are all obtained in time

$$\tau_4 = 2m(p-1),$$

and hence the recurrence $z = \hat{L}\hat{z} + g$ can be solved in time

$$\tau = \tau_3 + \tau_4 = \max \begin{cases} (2m^2+3m)p - \dfrac{m}{2}(2m^2+3m+5), \\ 2m(m+1)p - 2m. \end{cases}$$

Partition the recurrence $x = (I+L)x + f$ of n equations and bandwidth $m+1$ into the form

$$\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_k \end{bmatrix} = \begin{bmatrix} I+V_0 & & & \\ U_1 & I+V_1 & & \\ & \ddots & \ddots & \\ & & U_k & I+V_k \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_k \end{bmatrix} + \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_k \end{bmatrix},$$

where $I+V_0$ is of order m, and $I+V_i$, $i>0$ (except possibly for $I+V_k$) is of order s, i.e., $k = \lceil (n-m)/s \rceil$, and each U_i is of the form (19). Then, the solution vector x is given by

$$x_0 = (I+V_0)^* f_0, \tag{23}$$

$$x_i = (I+V_i)^*(f_i + U_i x_{i-1}), \qquad i = 1, 2, \cdots, k. \tag{24}$$

Equation (23) can be evaluated in $2(m-1)$ steps using the column sweep algorithm since $p > m$. The k equations in (24) can be solved one at a time using the algorithm developed for solving recurrence (18). Hence using p processors, the recurrence (12) of order n and bandwidth $m = 1$ can be solved in time

$$T_p = 2(m-1) + \tau\left\lceil \frac{(n-m)}{p(p+m-1)} \right\rceil,$$

where τ is given by equation (17). □

4. Linear systems $x = Ax + b$. In this section, we will deal with the general, linear system

$$x = Ax + b. \tag{25}$$

In § 1.3, we saw that the solution of the shortest path problem can be obtained by solving the system (25). Due to the form of (25), the n by n matrix A has, in

general, its only null elements on the main diagonal. The matrix A is called sparse if it has a small proportion of nonnull elements. If we let n_A be the number of nonnull elements of A, then we say that A is sparse if $n_A \ll n^2$.

We reiterate the assumption that the matrix A is definite, that is, the associated graph G has no cycles the sum of whose arc weights are less than or equal to zero. Now we present two direct methods and two iterative methods for solving the system (25). Since the Matrix A is definite, the iterative methods can also be posed as direct methods. We also discuss the special case when all of the arc weights of the underlying graph G are positive. The section is concluded with a comparison of the methods when the matrix A is sparse.

4.1. Elimination methods (Carré [2, § 7.2, 7.3]). The method of generalized Gaussian elimination was first proposed by Carré [2], [3]. We state the method and give time and processor counts for a parallel implementation.

ALGORITHM 5. *Parallel Gaussian elimination.*
For $k = 1, 2, \cdots, n-1$,

$$\left.\begin{aligned} \text{Set } \alpha_{i,j} &= \alpha_{i,k} \times \alpha_{k,j} + \alpha_{i,j} \\ \beta_i &= \alpha_{i,k} \times \beta_k + \beta_i \end{aligned}\right\} \quad \text{in parallel for } i, j = k+1, k+2, \cdots, n;\ i \neq j.$$

The resulting upper triangular system can be solved using any method of § 2. The reduction of the matrix A to upper triangular form can be accomplished in $2(n-1)$ steps using at most $(n-1)^2$ processors.

The added time and perhaps, added processors necessary to solve the resulting linear recurrence of Gaussian elimination can be avoided, by using the generalized Jordan elimination method of Carré [2, § 7.3].

ALGORITHM 6. *Parallel Jordan elimination.*
For $k = 1,2, \cdots, n-1$,

$$\left.\begin{aligned} \text{Set } \alpha_{i,j} &= \alpha_{i,k} \times \alpha_{k,j} + \alpha_{i,j} \\ \beta_i &= \alpha_{i,k} \times \beta_k + \beta_i \end{aligned}\right\} \quad \begin{aligned} &\text{in parallel for } i = 1, \cdots, n; \\ &j = k+1, k+2, \cdots, n;\ i \neq j. \end{aligned}$$

Note that upon termination, the vector b contains the solution x. Clearly, Algorithm 6 can be performed in $2(n-1)$ steps using at most n^2+1 processors. Thus, we see that by using an additional $2n$ processors over Gaussian elimination, we can solve the system (25) in the same time as it takes to do the elimination step of Gaussian elimination.

4.2. Iterative methods. In this section we present two iterative methods for solving the system (25). They correspond to the Jacobi and Gauss-Seidel methods for the iterative solution of linear algebraic equations. The methods of Bellman [1] and Moore [21] are sequential versions of the generalized Jacobi method. Ford and Fulkerson's [10] sequence of scanning the arcs in Ford's method [9] is a sequential version of the generalized Gauss–Seidel method.

The system (25) can be solved using the generalized Jacobi method

$$x^{(k)} = Ax^{(k-1)} + b. \tag{26}$$

Since the graph G associated with the distance matrix A is assumed to be definite, from Carré [2, Thm. 6.1], with an initial guess of a null $x^{(0)}$, we see that $x^{(n)} = x$.

We observe that the matrix-vector product $Ax^{(k-1)}$ can be formed in parallel in $\log n + 1$ steps using at most n^2 processors. Thus a single iteration requires $\log n + 2$

steps and at most n^2 processors. The overall algorithm can be accomplished in no more than $n(\log n + 2)$ steps using n^2 processors.

In our context, the generalized Jacobi method can be considered as a direct method. Since $x^{(n)}$ is the solution to the system (25), we can write it, by repeatedly using (26), as

$$x^{(n)} = A^n x^{(0)} + A^{n-1}b + A^{n-2}b + \cdots + Ab + b. \tag{27}$$

Since $x^{(0)}$ was chosen to be null, (27) simplifies to

$$x^{(n)} = A^{n-1}b + A^{n-2}b + \cdots + Ab + b, \tag{28}$$

which can be evaluated in parallel using the following algorithm.

ALGORITHM 7. *Direct Jacobi method.*
Set $x_0 = b$.
For $i = 0, 1, \cdots, \nu = \log(n/4)$
$x_{i+1} = (I + A^{2^i})x_i$ (we can also terminate whenever $x_{i+1} = x_i$)
$A^{2^{i+1}} = A^{2^i}A^{2^i}$
$x = (I + A^{n/2})x_{\nu+1}$ (if necessary).

This is the method proposed by Dekel and Sahni [6], for cube connected and perfect shuffle computers. By summing the time counts for Algorithm 7, and observing that the maximum number of processors occurs during the formation of $A^{2^{i+1}}$, we get the following theorem.

THEOREM 5. *The solution to the system* (25) *can be obtained in no more than* $2\log^2 n + 2\log n - 1$ *steps using* n^3 *processors.*

We will now discuss a parallel implementation of the generalized Gauss–Seidel method. We begin by observing that the matrix A of the system (25) can be written as

$$A = L + U. \tag{29}$$

Consequently, (25) becomes

$$x = (L + U)x + b, \tag{30}$$

which leads to the generalized Gauss–Seidel iteration

$$x^{(k+1)} = Lx^{(k+1)} + Ux^{(k)} + b.$$

Solving for $x^{(k+1)}$ yields

$$x^{(k+1)} = (I + L)^* Ux^{(k)} + (I + L)^* b \equiv Mx^{(k)} + c. \tag{31}$$

From Sameh and Brent [22, (2.16)], we get that the iteration matrix M and vector c can be formed in $\frac{1}{2}\log^2 n + O(\log n)$ steps using at most $n^3/6 + O(n^2)$ processors. Thus, the solution x can be found by the generalized Gauss–Seidel method in $n \log n + O(n)$ steps using at most $n^3/6 + O(n^2)$ processors.

Carré [2] has shown that the number of iterations required by the generalized Gauss–Seidel method is not greater than the number required by the generalized Jacobi method. It is thus likely that for many problems, the overall work for the generalized Gauss–Seidel method will be less than that of the generalized Jacobi method, when used in an iterative fashion. This potential saving in time has a substantial cost in additional processors.

When used as a direct method, the generalized Gauss–Seidel scheme would use Algorithm 7 applied to the matrix M and vector c of (31). Note that an additional

$\frac{1}{2}\log^2 n + O(\log n)$ operations are required to form the matrix M and vector c. Hence, for the generalized Gauss–Seidel method to be more efficient than the generalized Jacobi method, Algorithm 7 applied to the matrix M and vector c must take at least 25% fewer iterations than the same algorithm applied to the original matrix A and vector b. Consequently, one would prefer the generalized Jacobi method for solving the system (25).

4.3. Dijkstra's method. We now consider the special case when all of the arc weights of the associated graph G are positive. In the regular algebra this means that the elements of the matrix A of the system (25) satisfy $\alpha_{i,j} > \varepsilon$ for all i and j. Due to this added assumption, it is easy to see that there an be no linear algebraic analogues for algorithms tailored to these problems.

Dijkstra's method [7] is such an algorithm, iterative in nature with no linear algebraic analogue. In Dijkstra's method, nodes of the graph G are separated into two classes, temporary and permanent. All nodes start out temporary and become permanent one at a time. The algorithm terminates once all nodes have become permanent.

Let mode [*] be an array of bits used to indicate whether a given node is temporary or permanent. When mode $[i] = 0$, then node i is temporary. If mode $[i] = 1$, node i is permanent. If we wish to solve the system (25), where $\alpha_{i,j} > \varepsilon$ for all i and j, then Dijkstra's algorithm can be stated as follows.

ALGORITHM 8. *Parallel Dijkstra's algorithm.*
The vector b is overwritten by the solution x.
Initialize: mode $[i] = 0$, $i = 1, 2, \cdots, n$.
For $i = 1, 2, \cdots, n = 1$

(32) Find j such that $\beta_j = \sum_{\text{mode}[k]=0} \beta_k$,

(33) mode $[j] = 1$.

For all k such that mode $[k] = 0$

(34) $\beta_k = \beta_k + \beta_j \times \alpha_{j,k}$.

We observe that for each iteration, line (32) can be done in $\log(n+1-i)$ steps using $n-i$ processors. Overall, line (32) requires

$$\sum_{i=2}^{n} \log i = n \log \frac{n}{2}$$

operations and at most $n-1$ processors. Similarly line (33) requires one operation each iteration, and line (34) can be performed in parallel at each iteration using two operations and $n-i$ processors. The overall work for a parallel version of Dijkstra's method is $n \log n + 2n - 3$ steps using at most $n-1$ processors. Compared with n iterations of the generalized Jacobi method, Dijkstra's method requires essentially the same time, but far fewer processors ($n-1$ vs. n^2). Thus, for dense matrices A, Dijkstra's method appears to be the iterative method of choice.

4.4. Sparse matrix considerations. When the matrix A of the system (25) is large and sparse, that is the number of arcs n_A in the graph G is much less than n^2, direct methods are impractical. Unless the matrix A is banded or has a special structure, undesirable fill-in of the matrix A will occur. Since the economization of storage for large sparse matrices is essential, iterative methods must be used.

Not all iterative methods are appropriate for this problem. The generalized Gauss–Seidel method requires the formation of the iteration matrix $M = (I+L)^*U$. This matrix in general will be dense and hence too costly to store. We will thus restrict ourselves to a discussion of the generalized Jacobi method and Dijkstra's method.

Let q be the maximum number of nonnull elements in any row of A. In order to compare the two algorithms, we will use a cost function which is the product of the number of processors and time. For dense matrices, we see that Dijkstra's method has a cost of

$$(n-1)(n \log n + 2n - 3) = n^2 \log n + O(n^2).$$

Similarly, the generalized Jacobi method has a cost of $n^3 \log n + O(n^3)$ which is about n times greater than that of Dijkstra's method.

For sparse matrices, the generalized Jacobi method requires $n \log q + 2n$ steps and $n_A - n < n(q-1)$ processors. Dijkstra's method still requires $n \log n + 2n - 3$ steps and $n-1$ processors. So, the cost for the generalized Jacobi method becomes

$$n(q-1)(n \log q + 2n) = (q-1)n^2 \log q + O(qn^2),$$

while the cost of Dijkstra's method remains the same. When the generalized Jacobi method requires the full n iterations, it is more cost effective than Dijkstra's method if $(q-1)\log q < \log n$. In general, the generalized Jacobi method requires a fraction of n iterations. Let k represent the number of Jacobi iterations required for a given problem. Then if $k(q-1)\log q < n \log n$, the generalized Jacobi method is the superior algorithm. A similar tradeoff between Dijkstra's method and a version of Ford's method was found by Hulme and Wisniewski [14] for sequential algorithms.

REFERENCES

[1] R. BELLMAN, *On a routing problem*, Quart. Appl. Math., 16 (1958), pp. 87–90.

[2] B. A. CARRÉ, *An algebra for network routing problems*, J. Inst. Math. Appl., 7 (1971), pp. 273–294.

[3] ———, *An elimination method for minima-cost network flow problems*, in Large Sparse Sets of Linea· Equations, J. K. Reid, Ed., Proc. I.M.A. Conference, Oxford, 1970, Academic Press, London, 1971.

[4] Y. K. CHEN AND T. FENG, *A parallel algorithm for the maximum flow problem*, (summary), Proc. 1973 Sagmore Conf. on Parallel Processing (1973), p. 60.

[5] R. CRUON AND P. HERVÉ, *Quelques résultats relatifs à une structure algébrique et son application au problème central de l'ordonnancement*, Revue fr. Rech. opér., 34 (1965), 3–19.

[6] E. DEKEL AND S. SAHNI, *Parallel matrix and graph algorithms*, Tech. Rep. 79–10, Dept. Computer Science, University of Minnesota, June 1979.

[7] E. W. DIJKSTRA, *A note on two problems in connexion with graphs*, Numer. Math., 1 (1977), pp. 269–271.

[8] S. E. DREYFUS, *An appraisal of some shortest-path algorithms*, Operations Res., 17 (1969), pp. 395–412.

[9] L. R. FORD, Jr., *Network flow theory*, P-928, The Rand Corporation, Santa Monica, CA, August 1956.

[10] L. R. FORD, JR., AND D. R. FULKERSON, *Flows in networks*, Princeton University Press, Princeton, NJ, 1962.

[11] D. HELLER, *On the efficient computation of recurrence relations*, ICASE, Hampton, VA; Dept. Computer Science, Carnegie–Mellon University, Pittsburgh, PA, 1974.

[12] T. C. HU, *Revised matrix algorithms for shortest paths*, SIAM J. Appl. Math., 15 (1967), pp. 207–218.

[13] B. L. HULME, *MINDPT: a code for minimizing detection probability up to a given time away from a sabatoge target*, Rep. SAND 77-2039, Sandia Laboratories, Albuquerque, NM, Dec. 1977.

[14] B. L. HULME AND J. A. WISNIEWSKI, *A comparison of shortest path algorithms applied to sparse graphs*, SAND78-1411, Sandia Laboratories, Albuquerque, NM, August 1978.

[15] L. HYAFIL AND H. T. KUNG, *Parallel algorithms for solving triangular linear systems with small parallelism*, Dept. of Computer Science, Carnegie–Mellon University, Pittsburgh, PA, 1974.

[16] R. KALABA, *Proc. 10th Symposium on Applied Mathematics*, American Mathematical Society, Providence, RI, 1958, pp 1–280.

[17] D. J. KUCK, personal communication.
[18] D. J. KUCK AND A. H. SAMEH, *The Structure of Computers and Computation*, Vol. II, John Wiley, New York, in preparation.
[19] K. N. LEVITT AND W. H. KAUTZ, *Cellular arrays for the solution of graph problems*, Comm. ACM 15 (1972), pp. 789–801.
[20] G. C. MOISIL, *Asupra unor reprezentari ale grafurilor ce intervin in probleme de economia transporturilor*, Communicările Acad. Republicii Populare Romîne, 10 (1960), pp. 647–652.
[21] E. F. MOORE, *The shortest path through a maze*, in Proc. Internat. Symposium on Theory of Switching, Part II, Annals of the Computation Laboratory of Harvard University, Harvard University Press, Cambridge, MA, 1959, pp. 285–292.
[22] A. H. SAMEH AND R. P. BRENT, *Solving triangular systems on a parallel computer*, SIAM J. Numer. Anal., 14 (1977), pp. 1101–1113.
[23] G. B. VARNADO, et al., *Reactor safeguards system assessment and design, Volume I*, Rep. SAND77-0644, Sandia Laboratories, Albuquerque, NM, June 1978.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 3, September 1982

0196-5212/82/0303-0014 $01.00/0

ON THE EXPONENT OF A PRIMITIVE, NEARLY REDUCIBLE MATRIX. II*

JEFFREY A. ROSS†

Abstract. A nonnegative matrix is called nearly reducible provided it is irreducible and the replacement of any positive entry by zero yields a reducible matrix. The purpose of this article is to investigate the exponent $\gamma(A)$ of an $n \times n$ primitive, nearly reducible matrix A. Our principal result is that $\gamma(A) \leqq n + s(n-3)$, where s is the length of a shortest circuit in the directed graph associated with A. It is an easy application of this result to find gaps in the exponent set of $n \times n$ primitive, nearly reducible matrices. We also show that for integers n, k satisfying $n \geqq k-1 \geqq 5$ there exists an $n \times n$ primitive, nearly reducible matrix with exponent k. The proofs are carried out by means of directed graphs.

Key words. exponent, primitive nearly reducible matrix, minimally strong directed graph

1. Introduction. Let A be an $n \times n$ nonnegative matrix. Then A is *reducible* provided there exists a permutation matrix P such that

$$PAP^t = \begin{bmatrix} B & 0 \\ C & D \end{bmatrix}$$

where B and D are square (nonvacuous) matrices. The matrix A is *irreducible* if it is not reducible. The importance of irreducible matrices in the study of nonnegative matrices, especially the Perron–Frobenius theory, is well known [2], [13]. If A is an irreducible matrix, but any matrix obtained from A by changing a positive entry to zero is reducible, then A is said to be *nearly reducible.* Thus the nearly reducible matrices are the minimal irreducible matrices, and for this reason it is natural to investigate their properties [3].

Let A be an $n \times n$ irreducible matrix. The number $h(A)$ of eigenvalues of A of maximum modulus is called the *index of cyclicity* of A. If $h(A) = 1$, then A is said to be *primitive.* The index of cyclicity of A is dependent only on the zero-nonzero pattern of A, and this is reflected in the following equivalent condition for primitivity [13, p. 41–42]: The matrix A is primitive if and only if there exists a positive integer k such that A^k is a positive matrix. The least such integer k is called the *exponent* of A and is denoted $\gamma(A)$. It clearly suffices to consider only matrices of 0's and 1's when investigating primitive matrices and their exponents.

With each nonnegative matrix there is associated in a natural way a directed graph; this association and some properties of directed graphs will be given in § 2. The associated directed graph reflects the zero-nonzero pattern of the matrix, and determines some of the properties of the eigenvalues of the matrix. In particular, it determines the index of cyclicity and, in the case of a primitive matrix, the exponent.

Let A be an $n \times n$ primitive matrix of 0's and 1's. Then $\gamma(A) = 1$ if and only if A is positive. Wielandt [14] stated and Holladay and Varga [8] proved that $\gamma(A) \leqq$

* Received by the editors October 28, 1981, and in revised form February 24, 1982. This research was supported in part by the National Science Foundation under grant no. ISP-8011451 and a grant from the University of South Carolina Research and Productive Scholarship Fund. This paper was presented at the 1982 SIAM Conference on Applied Linear Algebra, Raleigh, NC, April 26–29, 1982.

† Department of Mathematics and Statistics, University of South Carolina, Columbia, South Carolina 29208.

$(n-1)^2+1$, with equality if and only if there exists a permutation matrix P such that

$$PAP^t = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Dulmage and Mendelsohn [6], and later Denardo [5], proved that

$$\gamma(A) \leqq n + s(n-2), \tag{1.1}$$

where s is the length of a shortest circuit in the directed graph associated with A. Exponents of particular classes of primitive matrices have also been investigated. Lewin [11] gave a tight upper bound on $\gamma(A)$ when the matrix A is doubly stochastic. Brualdi and this author [4] showed that if A is nearly reducible, then $6 \leqq \gamma(A) \leqq n^2-4n+6$, and both the upper and lower bounds are attained for every integer $n \geqq 4$. In addition, matrices attaining the upper bound were characterized. Other bounds for $\gamma(A)$ can be found in [5], [6], [7], [8] and [10].

In this paper we use graph-theoretic means to investigate the exponent of a primitive, nearly reducible matrix, extending the work in [4]. Our main result is an improvement of (1.1) when the matrix A is both primitive and nearly reducible. In particular, we show that in this case $\gamma(A) \leqq n + s(n-3)$. In addition, matrices meeting this bound (which is not always attainable) are characterized. We also obtain results on the existence of gaps in the exponent set of primitive, nearly reducible $n \times n$ matrices.

2. Minimally strong digraphs. Let $D = (X, U)$ be a directed graph (digraph) whose set of vertices is X and whose set of arcs is U. For $x, y \in X$, a *path* π from x to y is a sequence $(x_0, x_1, \cdots, x_k)$ of vertices with $k \geqq 0$ such that $x = x_0$, $y = x_k$, and $(x_i, x_{i+1}) \in U$ for $i = 0, 1, \cdots, k-1$. The arcs (x_i, x_{i+1}), $i = 0, 1, \cdots, k-1$, are the *arcs of* π and need not be distinct; the *length* of π, denoted $l(\pi)$, is k. The *initial vertex* of π is x_0, the *terminal vertex* is x_k, and $x_1, \cdots, x_{k-1}$ are the *internal vertices* of π. A path $(x_i, x_{i+1}, \cdots, x_j)$ with $0 \leqq i \leqq j \leqq k$ is a *subpath* of π. If $k \geqq 1$ and $x_0 = x_k$, then π is a *circuit*; a circuit of length 1 is a *loop*. An arc of D which is not an arc of the circuit π but joins two vertices of π is called a *chord* of π. The *distance from x to y (in D)*, denoted $d_D(x, y)$, is the length of a shortest path from x to y (where $d_D(x, y) = \infty$ if there does not exist a path from x to y). A path $(x_0, x_1, \cdots, x_k)$ is *elementary* provided $x_i \neq x_j$ for $0 \leqq i < j \leqq k$. A circuit $(x_0, x_1, \cdots, x_{k-1}, x_0)$ is *elementary* provided $x_i \neq x_j$ for $0 \leqq i < j \leqq k-1$. The digraph D is said to be *strongly connected* (or *strong*) if for each pair of vertices x, y there is a path from x to y and a path from y to x. A strong digraph D is *minimally strong* if each digraph obtained from D by the removal of an arc is not strongly connected. We denote by D^k the digraph with vertex set X and arc set $U^k = \{(x, y)$: there exists in D a path of length k from x to $y\}$.

Let $A = [a_{ij}]$ be a nonnegative $n \times n$ matrix. Let $X_A = \{x_1, \cdots, x_n\}$ and define $U_A = \{(x_i, x_j)$: $a_{ij} \neq 0$, $1 \leqq i, j \leqq n\}$. The digraph $D(A) = (X_A, U_A)$ is the *digraph associated with A*. Note that given a digraph $D = (X, U)$ on n vertices $x_1, \cdots, x_n$, we may associate with D an $n \times n$ matrix $A(D) = [a_{ij}^D]$ defined by $a_{ij}^D = 1$ if and only if $(x_i, x_j) \in U$. Then $D(A(D)) \cong D$. It is well known [13, p. 20] that A is irreducible if and only if $D(A)$ is strong. It follows that A is nearly reducible if and only if $D(A)$ is minimally strong. Let d be the greatest common divisor of the lengths of the

(elementary) circuits of $D(A)$. Then [13, p. 49–50] $h(A) = d$. Thus, we say a digraph D has *index of cyclicity* $h(D)$ equal to the greatest common divisor of the lengths of its elementary circuits, and D is *primitive* if $h(D) = 1$. We also define the *exponent* of a primitive digraph D, denoted $\gamma(D)$, to be the exponent of $A(D)$. Letting $A(D)^k = a_{ij}^{(k)}$, we have $a_{ij}^{(k)} > 0$ if and only if there exists an arc from x_i to x_j in D^k. Thus $\gamma(D)$ is equal to the least integer k such that for any ordered pair of (not necessarily distinct) vertices x, y there exists in D a path from x to y of length k.

We now briefly discuss properties of minimally strong directed graphs. We refer the reader to [1] and [12] for further details. The following properties are readily established.

LEMMA 2.1. *Let $D = (X, U)$ be a minimally strong digraph and let $Y \subseteq X$. Then the following properties hold*:

(2.1) *D has no loops.*

(2.2) *If the digraph D_Y induced on Y is strong, then D_Y is minimally strong.*

(2.3) *A circuit has no chords.*

(2.4) *If $(x, y) \in U$, then (x, y) is an arc of every path from x to y.*

In a strong digraph D each vertex x has indegree $\delta_D^-(x)$ and outdegree $\delta_D^+(x)$ at least one. A vertex x is called an *antinode* if $\delta_D^+(x) = \delta_D^-(x) = 1$; otherwise, x is called a *node*. Suppose D is minimally strong and $\pi = (x_0, x_1, \cdots, x_k)$ is a path with $k \geqq 2$ whose initial and terminal vertices are nodes and whose internal vertices are antinodes. Letting $Y = \{x_1, \cdots, x_{k-1}\}$, we say π is a *branch* provided D_{X-Y} is strong (in [12] we call this a superfluous branch; here all branches are superfluous). By (2.2) of Lemma 2.1 we know that D_{X-Y} is minimally strong, and we write $D_{X-Y} = D \sim \pi$. We say $D \sim \pi$ is formed from D by *removing the branch* π, or D is formed from $D \sim \pi$ by *adding the branch* π. Every minimally strong digraph which is not an elementary circuit contains a branch [1, p. 32]. In fact, the following stronger result is proved in [12].

LEMMA 2.2. *Let D be a minimally strong digraph which is not an elementary circuit, and let ρ be a branch of D. Then there exists a branch π of D such that ρ and π do not lie on a common elementary circuit of D.*

COROLLARY 2.3. *Let D be a minimally strong digraph which is not an elementary circuit, and suppose D contains an elementary circuit of length s. Then there exists a branch π of D such that $D \sim \pi$ contains an elementary circuit of length s.*

Let D be a minimally strong digraph. If a branch is removed from D, the resulting minimally strong digraph either is an elementary circuit or contains a branch. Thus one may continue to remove branches from the resulting digraphs until an elementary circuit is obtained. The number of branches which must be removed to obtain an elementary circuit is an invariant of the digraph ([1], [12]) and is denoted by $\mu(D)$.

Now let D be a minimally strong digraph which is not an elementary circuit. Suppose there exists a sequence $D_0, D_1, \cdots, D_\mu = D$ of minimally strong digraphs which satisfy

(2.5) D_0 is an elementary circuit.

(2.6) D_i is formed from D_{i-1} $(i = 1, \cdots, \mu)$ by adding the branch $\pi_i = (x_0^i, x_1^i, \cdots, x_{k_i}^i)$ such that for $i = 2, \cdots, \mu$, we have $x_0^i = x_{r_i}^{i-1}$ and $x_{k_i}^i = x_{s_i}^{i-1}$ where $1 \leqq s_i \leqq r_i \leqq k_{i-1} - 1$.

In this case we say the digraph D is *special.* Note that for $i = 1, \cdots, \mu - 1$, the path π_i is a branch of D_i but not a branch of D. Also, notice that $\mu = \mu(D)$.

For a minimally strong digraph D, let $\mathcal{B}(D)$ denote the number of branches of D. The proofs of the following two lemmas can be found in [12].

LEMMA 2.4. *Let D be a minimally strong digraph which is not an elementary circuit, and let $\pi = (x_0, x_1, \cdots, x_k)$ be a branch of D. Then either every branch of $D \sim \pi$ contains a subpath which is a branch of D, or there exists a branch $\rho = (y_0, y_1, \cdots, y_l)$ of $D \sim \pi$ such that no subpath of ρ is a branch of D but every other branch of $D \sim \pi$ is a branch of D. In the latter case $x_0 = y_i$ and $x_k = y_j$ where $1 \leqq j \leqq i \leqq l - 1$.*

LEMMA 2.5. *Let D be a minimally strong digraph which is not an elementary circuit. Then $\mathcal{B}(D) \geqq 2$, with equality if and only if D is special.*

Now let $D = (X, U)$ be an arbitrary digraph, and let $Y \subseteq X$ with $Y \neq \varnothing$. Let $y \notin X$ and form the digraph $D * Y = ((X - Y) \cup \{y\}, U_Y)$, where $(u, v) \in U_Y$ if and only if one of the following holds.

(2.7) $u, v \in X - Y$ and $(u, v) \in U$.

(2.8) $u \in X - Y$, $v = y$, and there exists $w \in Y$ such that $(u, w) \in U$.

(2.9) $u = y$, $v \in X - Y$, and there exists $w \in Y$ such that $(w, v) \in U$.

We say $D * Y$ is the *contraction* of Y *(in D).* The following fact is easily proved.

LEMMA 2.6. *Let $D = (X, U)$ be a strong digraph, and let $Y \subseteq X$ with $Y \neq \varnothing$. Then $D * Y$ is strong.*

3. Small exponents. In [4] it is shown that if D is a primitive, minimally strong digraph, then $\gamma(D) \geqq 6$, and for each $n \geqq 4$ there exists a primitive, minimally strong digraph on n vertices with exponent six. Here we show that for $n \geqq 5$, each of the integers 6 through $n + 1$ may be achieved as the exponent of a primitive, minimally strong digraph on n vertices.

Let $D = (X, U)$ be a primitive digraph, and let $x \in X$. Then there exists an integer f with the property that for every vertex $y \in X$ there exists a path from x to y of length f. The least such integer f is called the *reach* of x and is denoted by $f_D(x)$. The following two lemmas can both be found in [6].

LEMMA 3.1. *Let $D = (X, U)$ be a primitive digraph and let $x \in X$. If $p \geqq f_D(x)$, then for any $y \in X$ there exists a path of length p from x to y.*

LEMMA 3.2. *Let $D = (X, U)$ be a primitive digraph. Then* $\gamma(D) = \max\{f_D(x): x \in X\}$.

We now are ready for the main theorem of this section.

THEOREM 3.3. *Given integers $n \geqq 5$ and $k \geqq 6$ with $k \leqq n + 1$, there exists a primitive, minimally strong digraph $G_{n,k}$ on n vertices with $\gamma(G_{n,k}) = k$.*

Proof. We first show, using induction on k, the existence of the primitive minimally strong digraph $G_{k-1,k}$ on $k - 1$ vertices with $\gamma(G_{k-1,k}) = k$. Then for each $n \geqq k - 1$ we will be able to add $n - k + 1$ vertices and $2(n - k + 1)$ arcs to $G_{k-1,k}$ in a way to obtain a primitive, minimally strong digraph $G_{n,k}$ with $\gamma(G_{n,k}) = \gamma(G_{k-1,k}) = k$.

It is easily verified that the digraph $G_{5,6}$ of Fig. 1 is a primitive, minimally strong digraph on 5 vertices with $\gamma(G_{5,6}) = 6$.

Suppose $k = 2i + 2$ is even ($i \geqq 2$). Let $G_i = (X_i, U_i)$ be the digraph of Fig. 2. Then G_i is primitive and minimally strong. We have $G_{5,6} \cong G_2$, so by the inductive hypothesis we may assume that $\gamma(G_i) = 2i + 2 = k$.

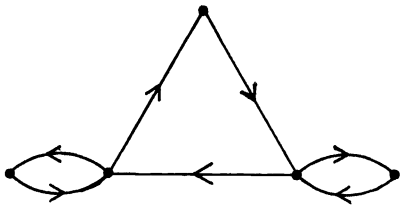


FIG. 1

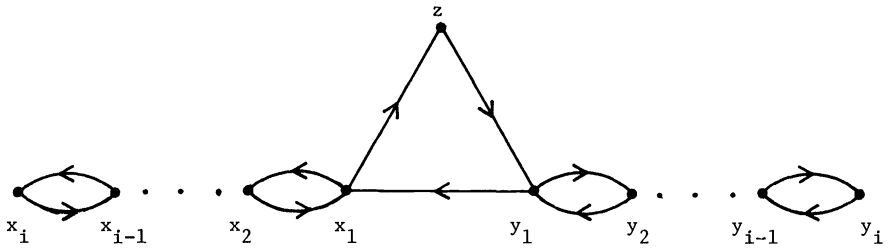


FIG. 2. *The digraph* G_i.

Let the digraphs $E_i = (Y_i, V_i)$ and $F_i(Z_i, W_i)$ be those illistrated in Fig. 3 ($i \geqq 2$). Thus we have

$$Y_i = X_i \cup \{x_{i+1}\}, \qquad V_i = U_i \cup \{(x_i, x_{i+1}), (x_{i+1}, x_i)\}, \quad \text{and}$$

$$Z_i = X_i \cup \{y_{i+1}\}, \qquad W_i = U_i \cup \{(y_i, y_{i+1}), (y_{i+1}, y_i)\}.$$

Both E_i and F_i are primitive and minimally strong. We show by induction that $\gamma(E_i) = \gamma(F_i) = k+1$.

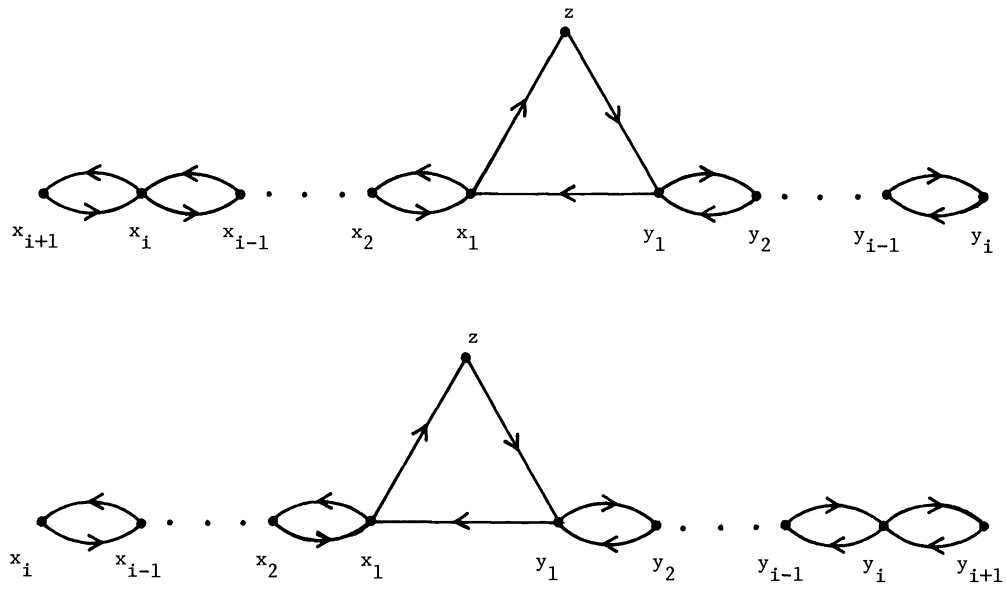


FIG. 3. *The digraphs* E_i *(top) and* F_i *(bottom).*

In E_i there exists only one elementary path from x_{i+1} to y_i, and this path has length $2i+1=k-1$. There is no path in E_i from x_{i+1} to y_i of length k, and hence by Lemma 3.1, $f_{E_i}(x_{i+1}) \geqq k+1$. From applying Lemma 3.2 to G_i and using the inductive hypothesis, it follows that there exists a path of length k in G_i, hence in E_i, from x_i to any vertex of X_i. Thus there exists a path of length $k+1$ from x_{i+1} to any vertex of X_i. Also, the path $(x_{i+1}, x_i, \cdots, x_1, z, y_1, x_1, x_2, \cdots, x_{i+1})$ is a path of length $2i+3=k+1$ from x_{i+1} to x_{i+1}, so $f_{E_i}(x_{i+1})=k+1$. Now let $x \in X_i$. It follows from applying Lemmas 3.1 and 3.2 to G_i that there exist paths of length k and of length $k+1$ in G_i, hence in E_i, from x to any vertex of X_i. Since there exists a path from x to x_i of length k, there exists a path of length $k+1$ from x to x_{i+1}. Thus for each $x \in X_i$, we have $f_{E_i}(x) \leqq k+1$. Hence it now follows from Lemma 3.2 that $\gamma(E_i)=k+1$. In a similar manner one may show that $f_{F_i}(x_i)=k+1$ and $\gamma(F_i)=k+1$.

We now continue the induction to show that $\gamma(G_{i+1})=k+2=2(i+1)+2$. The only elementary path in G_{i+1} from x_{i+1} to y_{i+1} has length $2i+2=k$, and there is no path of length $k+1$ from x_{i+1} to y_{i+1} in G_{i+1}. Hence $f_{G_{i+1}}(x_{i+1}) \geqq k+2$. Using arguments similar to those above, since $\gamma(E_i)=\gamma(F_i)=k+1$, we have that $f_{G_{i+1}}(x_{i+1})=k+2$ and $f_{G_{i+1}}(x) \leqq k+2$ for any other vertex x of G_{i+1}. It now follows from Lemma 3.2 that $\gamma(G_{i+1})=k+2$. Thus by taking $G_{k-1,k}=G_i$ for $k=2i+2$ and $G_{k-1,k}=E_i$ for $k=2i+3$, we have that $G_{k-1,k}$ is a primitive, minimally strong digraph on $k-1$ vertices with $\gamma(G_{k-1,k})=k$.

Let $i \geqq 2$. For $k=2i+2$ and $n \geqq k$, let $G_{n,k}=(X_i \cup X'_{n-k+1}, U_i \cup U'_{n-k+1})$, and for $k=2i+3$ and $n \geqq k$, let $G_{n,k}=(Y_i \cup X'_{n-k+1}, V_i \cup U'_{n-k+1})$, where

$$X'_{n-k+1}=\{x'_1, x'_2, \cdots, x'_{n-k+1}\}, \quad \text{and} \quad U'_{n-k+1}=\{(x'_j, x_1), (x_1, x'_j): j=1, \cdots, n-k+1\}.$$

The digraph $G_{n,k}$ is illustrated in Fig. 4 (in which $l=[(k-1)/2]$). It is straightforward to verify that for $n \geqq k-1 \geqq 5$ the digraph $G_{n,k}$ is a primitive, minimally strong digraph whose exponent is k. □

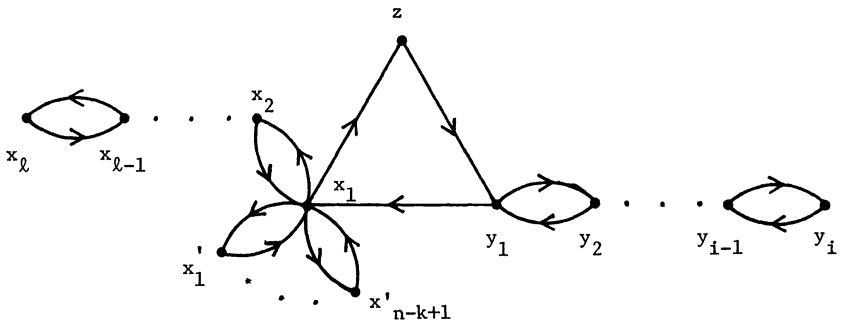


FIG. 4

COROLLARY 3.4. *Given integers $n \geqq 5$ and $k \geqq 6$ with $k \leqq n+1$, there exists an $n \times n$ primitive, nearly reducible matrix A with $\gamma(A)=k$.*

4. An upper bound. It is shown in [4] that if A is an $n \times n$ primitive, nearly reducible matrix, then $\gamma(A) \leqq n+(n-2)(n-3)$, and matrices for which equality holds are characterized. In this section we generalize this result, obtaining an upper bound on $\gamma(A)$ depending on the length of a shortest circuit of $D(A)$. Let $s(D)$ denote the length of a shortest circuit in a strong digraph D, and for integers a, b, let (a, b) denote their greatest common divisor.

THEOREM 4.1. *Let A be an $n \times n$ primitive, nearly reducible matrix of 0's and 1's, and let $s = s(D(A))$. Then*

$$\gamma(A) \leqq n + s(n-3), \tag{4.1}$$

with equality if and only if there exists a permutation matrix P such that

$$PAP^t = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \cdots & 1 \\ 0 & 1 & 0 & \cdots & 0 & \cdots & 0 \end{bmatrix},$$

where row s of PAP^t has the entry 1 in the first column. In particular, if $(s, n-1) \neq 1$, then $\gamma(A) < n + s(n-3)$, and if $(s, n-1) = 1$, then there exists an $n \times n$ primitive, nearly reducible matrix with exponent $n + s(n-3)$.

In order to prove Theorem 4.1, we shall first prove some useful lemmas concerning digraphs.

LEMMA 4.2. *Let D be a minimally strong digraph which is not an elementary circuit. Suppose there exists at most one branch π of D for which $s(D \sim \pi) = s(D)$. Then D contains exactly two branches, one a branch π for which $s(D \sim \pi) = s(D)$, the other a branch ρ for which $s(D \sim \rho) > s(D)$. In particular, D is special.*

Proof. The proof is by induction on $\mu = \mu(D)$. The result holds if $\mu = 1$ since all such digraphs are special. So assume $\mu > 1$. It follows from Corollary 2.3 that there exists a branch π of D such that $s(D \sim \pi) = s(D)$. Let $D' = D \sim \pi$, and suppose that D' has two distinct branches π' and π'' such that $s(D' \sim \pi') = s(D')$ and $s(D' \sim \pi'') = s(D')$. It follows from Lemma 2.4 that some subpath τ of π' or π'' is a branch of D. However, since $s(D') = s(D)$, it now follows that $s(D \sim \tau) = s(D)$. Since $\tau \neq \pi$, this is a contradiction. Now $\mu(D') = \mu(D) - 1$, so by induction D' has exactly two branches, a branch α for which $s(D' \sim \alpha) = s(D')$ and a branch ρ for which $s(D' \sim \rho) > s(D')$. Also, D' is special. Since no subpath of α can be a branch of D, it follows that ρ and π are the branches of D, and we know $s(D \sim \pi) = s(D)$ and $s(D \sim \rho) > s(D)$. Finally, since $\mathscr{B}(D) = 2$, by Lemma 2.5 we conclude that D is special. □

LEMMA 4.3. *Let D be a primitive, minimally strong digraph. Suppose that for any branch α of D, either $s(D \sim \alpha) > s(D)$ or every circuit in $D \sim \alpha$ has length divisible by $s(D)$. Then D is special, and its two branches π and ρ satisfy*

$$s(D \sim \rho) > s(D). \tag{4.2}$$

$$\text{Every circuit in } D \sim \pi \text{ has length divisible by } s(D). \tag{4.3}$$

Moreover, D contains a unique circuit of length $s(D)$.

Proof. It follows from Corollary 2.3 that there exists a branch $\pi = (x_0, x_1, \cdots, x_k)$ for which (4.3) holds. Now let σ be any path from x_k to x_0 in $D \sim \pi$, and suppose $l(\sigma) = c$. Then (π, σ) is a circuit of length $c + k$. Let $g = (c + k, s)$, where $s = s(D)$. Since π is a branch of D, there exists a path λ in $D \sim \pi$ from x_0 to x_k with, say, $l(\lambda) = j$. Then $s \mid j + c$ since (λ, σ) is a circuit of $D \sim \pi$.

Now let β be any elementary circuit of D. If π is not a subpath of β, then $g \mid l(\beta)$ since $s \mid l(\beta)$. Suppose that π is a subpath of β, so $\beta = (\pi, \tau)$ where τ is a path in $D \sim \pi$. Letting $l(\tau) = d$, we have $s \mid j + d$ since (λ, τ) is a circuit of $D \sim \pi$. Hence $s \mid d - c$, and

from this it follows that $g|d+k$, that is, $g|l(\beta)$. Thus every elementary circuit of D has length divisible by g, and we conclude that $g=1$ since D is primitive. It now follows that every elementary circuit of D of the form (π, τ) has length coprime to s.

Suppose now that there exists another branch $\pi' \neq \pi$ for which every circuit of $D \sim \pi'$ has length divisible by s. Since π is a path of $D \sim \pi'$, there exists in $D \sim \pi'$ an elementary circuit of the form (π, τ). But this circuit has length coprime to s, contrary to assumption, so π is the unique branch of D satisfying (4.3). By assumption, any other branch of α of D satisfies $s(D \sim \alpha) > s(D)$. It now follows from Lemma 4.2 that there exists a unique branch $\rho \neq \pi$ for which (4.2) holds, that ρ and π are precisely the branches of D, and D is special. Since D is special and contains a unique branch ρ for which $s(D \sim \rho) > s$, one now sees that D contains a unique circuit of length s. □

One should note that it is quite possible to find primitive, minimally strong digraphs which satisfy the hypotheses of Lemmas 4.2 and 4.3.

Let $D=(X, U)$ be a digraph and let $x, y \in X$. We say x and y are *connected* in D if there exists a sequence of vertices $(x_0, x_1, \cdots, x_k)$ with $k \geqq 0$ such that $x = x_0$, $y = x_k$, and either $(x_i, x_{i+1}) \in U$ or $(x_{i+1}, x_i) \in U$ for $i = 0, 1, \cdots, k-1$. This is clearly an equivalence relation, and the equivalence classes are the *connected components* of D.

A somewhat stronger form of the following lemma is stated and proved in terms of matrices in [13, p. 43].

LEMMA 4.4. *Let E be a strong digraph with index of cyclicity h. Then for each integer $j \geqq 1$, the digraph E^{hj} has h connected components. Moreover, each connected component is strongly connected and primitive. In particular, for every integer $j \geqq 1$, E^j is strongly connected and primitive whenever E is.*

The following lemma is evident.

LEMMA 4.5. *Let E be a strong digraph with index of cyclicity h. If $(x_0, x_1, \cdots, x_h)$ is any path in E, then for each integer $j \geqq 1$, the vertices $x_0, \cdots, x_{h-1}$ are all in different connected components of E^{hj}, and x_0, x_h are in the same component.*

LEMMA 4.6. *Let $D=(X, U)$ be a primitive digraph and let $x \in X$ be a vertex on a circuit of length $s = s(D)$. Let $k = \max\{d_D(y, x): y \in X\}$ and let $d = \max\{d_{D^s}(x, y): y \in X\}$. Then*

$$\gamma(D) \leqq k + sd. \tag{4.4}$$

Proof. The proof of this lemma closely follows the proof of Theorem 1 of [6], so we shall be brief. Let $y \in X$ and let $l_y = d_D(y, x)$, so $l_y \leqq k$. Since x is a loop vertex of D^s, there exists in D^s a path of length exactly d from x to any vertex in X. Hence in D there exists a path of length $l_y + sd$ from y to any vertex of D, so $f_D(y) \leqq l_y + sd \leqq k + sd$. The inequality (4.4) now follows from Lemma 3.2. □

LEMMA 4.7. *Let D be a primitive special digraph with $\mu(D) \geqq 2$ and let $s = s(D)$. Suppose the branches ρ and $\pi = (x_0, x_1, \cdots, x_k)$ of D satisfy (4.2) and (4.3). Then $(D \sim \pi)^s$ has s connected components $A_1, \cdots, A_s$ which are strongly connected. Let*

$$D^* = (\cdots((D^s * A_1) * A_2) \cdots * A_s). \tag{4.5}$$

If $s \geqq 3$ and D^* is not an elementary circuit, then $\gamma(D) \leqq n - 1 + s(n-3)$.

Proof. Since $D \sim \pi$ still contains a circuit of length s, it follows from (4.3) that $h(D \sim \pi) = s$. The first conclusion now follows from Lemma 4.4. Since D^s is strong, it follows from Lemma 2.6 that D^* is strong. By Lemma 4.5, each vertex on a circuit of length s is in a different component of $(D \sim \pi)^s$. Moreover, these are loop vertices

in $(D \sim \pi)^s$ since in $D \sim \pi$ there is a path of length s from each one to itself. For $i = 1, \cdots, s$, let $a_i \in A_i$ be a loop vertex.

Since D is special, let $D_0, D_1, \cdots, D_\mu = D$ be a sequence of minimally strong digraphs satisfying (2.5) and (2.6). This sequence can be chosen so that ρ is a path of D_0 and $\pi = \pi_\mu$. Since $\mu(D) \geqq 2$, $D \sim \pi = D_{\mu-1} \neq D_0$. Since D_0 is an elementary circuit of length s, which by Lemma 4.3 is the unique circuit of size s in D, it now follows that the elementary circuit of D_1 containing π_1 has at least $2s$ vertices, and hence that π_1 has at least $s+1$ internal vertices. Thus there exist s consecutive internal vertices of π_1 which by Lemma 4.5 are all in different components of $(D \sim \pi)^s$. Hence $|A_i| \geqq 2$ for every $i = 1, \cdots, s$.

It follows from Lemma 4.5 and the definition of $h(D)$ that for each $i = 1, \cdots, s$, there exists an integer t_i with $0 \leqq t_i < s$ such that $d_{D \sim \pi}(x, x_0) \equiv t_i \pmod{s}$ for every $x \in A_i$. Furthermore $\{t_i : i = 1, \cdots, s\}$ is a complete set of residues (mod s), and for each $i = 1, \cdots, s$, there exists a vertex $v_i \in A_i$ such that $d_{D \sim \pi}(v_i, x_0) = t_i$. For each $i = 1, \cdots, s$, let $\mathcal{Y}_i$ be the contraction of A_i in D^*. Since the length of the elementary circuit σ of D containing π is greater than s, it now follows that $\delta^+_{D^*}(\mathcal{Y}_i) = 1$.

Suppose now that D^* is not an elementary circuit. Since D^* is strong, there must exist a vertex x of D^* such that $\delta^+_{D^*}(x) \geqq 2$. We have seen that $x \neq \mathcal{Y}_i$ for any $i = 1, \cdots, s$, so x must be an internal vertex of π. Choose $\mathcal{Y}_p$ such that $d_{D^*}(\mathcal{Y}_p, x) = \min \{d_{D^*}(\mathcal{Y}_i, x) : i = 1, \cdots, s\}$. Again, $l(\sigma) > s$ implies that $\delta^+_{D^*}(x) = 2$, and the terminal vertices of these arcs are an internal vertex z of π and $\mathcal{Y}_r$ for some $r = 1, \cdots, s$. Since $(l(\sigma), s) = 1$, $r \neq p$. Now let q be such that $d_{D^*}(z, \mathcal{Y}_q) = \min \{d_{D^*}(z, \mathcal{Y}_i) : i = 1, \cdots, s\}$. This defines q uniquely; since $(l(\sigma), s) = 1$, $q \neq p$, and since we also have $s \geqq 3$, $q \neq r$. It now follows that in D^s there exists a path from a_p to any vertex which avoids either A_r or A_q. Since $|A_r| \geqq 2$ and $|A_q| \geqq 2$, it follows from Lemma 4.6 that $\gamma(D) \leqq n - 1 + s(n-3)$. □

Suppose $D = (X, U)$ is a primitive, minimally strong digraph with a branch π for which $h(D \sim \pi) = h > 1$. Let $A_1, \cdots, A_h$ be the components of $(D \sim \pi)^s$. A vertex $x \in A_i$ $(i = 1, \cdots, h)$ is an *exit vertex* if in D^s there is an arc from x to either an internal vertex of π or to a vertex in some A_j with $j \neq i$.

Let $D = (X, U)$ be a primitive digraph and let $p_1, \cdots, p_k$ be the distinct lengths of the elementary circuits of D. For each ordered pair of vertices x, y we define a nonnegative integer $r_{x,y}$ as follows: $r_{x,y}$ is the length of a shortest path from x to y which for each $i = 1, \cdots, k$ contains a vertex of some circuit of length p_i. Note that $r_{x,x} = 0$ if for each $i = 1, \cdots, k$, x is a vertex of an elementary circuit of length p_i. Let $r(D) = \max \{r_{x,y} : x, y \in X\}$. An ordered pair of vertices x, y has the *unique path property* if every path from x to y of length at least $r_{x,y}$ consists of some path π from x to y of length $r_{x,y}$ augmented by elementary circuits each of which has a vertex in common with π.

Now let $p_1, \cdots, p_k$ be relatively prime positive integers and let $F(p_1, \cdots, p_k)$ be the largest integer which cannot be expressed in the form $a_1 p_1 + \cdots + a_k p_k$ where $a_1, \cdots, a_k$ are nonnegative integers. It is well known [9, p. 6] that $F(p_1, \cdots, p_k)$ is finite and that when $k = 2$, $F(p_1, p_2) = p_1 p_2 - p_1 - p_2$. The following is due to Dulmage and Mendelsohn [6].

LEMMA 4.8. *Let D be a primitive digraph for which $p_1, \cdots, p_k$ are the distinct lengths of the elementary circuits. Then*

$$\gamma(D) \leqq F(p_1, \cdots, p_k) + r(D) + 1. \tag{4.6}$$

If the ordered pair of vertices x, y has the unique path property, then

$$F(p_1, \cdots, p_k) + r_{x,y} + 1 \leqq \gamma(D). \tag{4.7}$$

We are now prepared to prove the main theorem. We restate it and prove it in terms of digraphs.

THEOREM 4.9. *Let $D=(X, U)$ be a primitive, minimally strong digraph on n vertices, and let $s=s(D)$. Then*

$$\gamma(D)\leqq n+s(n-3), \tag{4.8}$$

with equality if and only if D is isomorphic to the digraph $D_{s,n}$ of Fig. 5. In particular, if $(s, n-1)\neq 1$, then $\gamma(D)<n+s(n-3)$, and if $(s, n-1)=1$, then $D_{s,n}$ is a primitive, minimally strong digraph on n vertices with exponent $n+s(n-3)$.

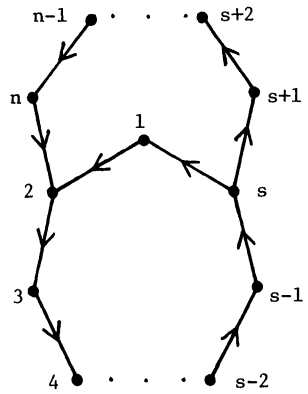


FIG. 5. *The digraph $D_{s,n}$.*

Proof. The proof is by induction on $\mu=\mu(D)$. If $\mu=1$, then D is isomorphic to a digraph of the type shown in Fig. 6, where we may assume $p<n-s$ so $s(D)=s$. Also, since D is minimally strong, $p\geqq 1$, and since D is primitive, $(s, n-p)=1$. Clearly $r(D)=r_{s+1,n}=n-p+n-s-1$. The vertices $s+1$ and n have the unique path property, so by Lemma 4.8,

$$\gamma(D)=F(n-p, s)+r(D)+1=n+s(n-2-p)\leqq n+s(n-3),$$

with equality if and only if $p=1$, if and only if $D\cong D_{s,n}$. This proves the theorem when $\mu=1$.

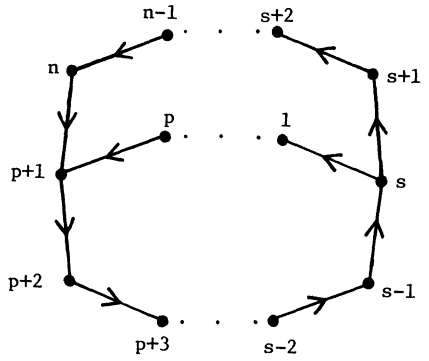


FIG. 6

Now suppose $\mu(D) \geqq 2$. We must show that $\gamma(D) \leqq n-1+s(n-3)$. Suppose first that D contains a branch π such that $s(D \sim \pi) = s(D)$ and $D \sim \pi$ is primitive. Let k equal the number of internal vertices of π. By the inductive hypothesis we have

$$\gamma(D \sim \pi) \leqq n-k+s(n-k-3) \leqq n+s(n-3)-3k. \tag{4.9}$$

Since $D \sim \pi$ has paths between any two vertices of all lengths $\gamma(D \sim \pi)$ and greater, it follows that $\gamma(D) \leqq \gamma(D \sim \pi)+2k$. Combining this with (4.9), we have $\gamma(D) \leqq n+s(n-3)-k \leqq n-1+s(n-3)$.

Hence we may assume that the removal of any branch π from D yields a digraph $D \sim \pi$ with either $s(D \sim \pi) > s(D)$ or $D \sim \pi$ not primitive. By Corollary 2.3 there exists a branch π with $s(D \sim \pi) = s(D)$ so we have $h(D \sim \pi) = h > 1$. First assume that $h \neq s$, so $h < s$ and $h | s$. By Lemma 4.4, $(D \sim \pi)^s$ has h connected components, each strongly connected and primitive. It follows from Lemma 4.5 that $(D \sim \pi)^s$ has at least s/h loop vertices in each component.

Suppose $s/h \geqq 3$. Let A be any connected component of $(D \sim \pi)^s$, and let $y \in A$ be an exit vertex of D^s. Let $x \in A$ be a loop vertex of D^s such that $d_{D^s}(x, y)$ is minimum. Since A contains at least two other loop vertices, it now follows that for any vertex $z \in X$, there exists in D^s a path from x to z of length at most, hence exactly, $n-3$. We may now apply Lemma 4.6 to conclude that $\gamma(D) \leqq n-1+s(n-3)$.

Suppose now that $s/h = 2$. If there exists more than one circuit of length s in $D \sim \pi$, then by Lemma 4.5 some component A of $(D \sim \pi)^s$ contains at least three loop vertices, and the result follows as above. So we may assume $D \sim \pi$ contains exactly one circuit of length s, and now by Lemma 4.5 each component of $(D \sim \pi)^s$ contains exactly two loop vertices. Let $\sigma = (x_0, x_1, \cdots, x_{s-1}, x_0)$ be the circuit of $D \sim \pi$ of length s. Since $\mu(D) \geqq 2$, there exists a vertex of σ, say x_{s-1}, such that $\delta^+_{D \sim \pi}(x_{s-1}) \geqq 2$. Thus there exists a vertex $x \neq x_0$ with (x_{s-1}, x) an arc of $D \sim \pi$. By (2.3) of Lemma 2.1 the vertex x is not a vertex of σ. Let A be the component of $(D \sim \pi)^s$ containing x_0, so we also have $x \in A$ and, by Lemma 4.5, $x_h \in A$. Since σ is the only circuit of $D \sim \pi$ of length s and $h = s/2$, it follows from (2.4) of Lemma 2.1 that the shortest path from x to a vertex of σ has length at least $s/2+1$. Hence in $D \sim \pi$ there exists a path of length s from x_h to $x' \in A$ where x_0, x_h, x, and x' are all distinct. In particular we can now conclude that if y is an exit vertex of A, either $d_{(D \sim \pi)^s}(x_0, y) \leqq |A|-3$ or $d_{(D \sim \pi)^s}(x_h, y) \leqq |A|-3$. Without loss of generality assume the former. In D there exists a path from any vertex z to x_0 of length at most $n-1$. Since x_0 is a loop vertex of D^s, and each component of $(D \sim \pi)^s$ contains at least two vertices, in D^s there is a path from x_0 to any vertex of length $n-3$. Therefore $\gamma(D) \leqq n-1+s(n-3)$ by Lemma 4.6.

Thus we are left only to consider the case when the removal of any branch π from D yields a digraph $D \sim \pi$ with either $s(D \sim \pi) > s(D)$ or $h(D \sim \pi) = s(D)$. In this case D satisfies the hypotheses of Lemma 4.3, so D is special, and its branches ρ and π satisfy (4.2) and (4.3). Moreover, D contains a unique (elementary) circuit of length $s = s(D)$. For the remainder of this proof we shall consider D to be such a digraph.

We first suppose $s \geqq 3$, and let $\pi = (x_0, \cdots, x_k)$ and $\tau = (y_0, \cdots, y_l)$, where τ is the branch of $D \sim \pi$ which is not ρ. By (2.6) we have $x_0 = y_j$ and $x_k = y_i$ where $1 \leqq i \leqq j \leqq l-1$. Let $\alpha = (y_i, \cdots, y_j)$ and let $\beta = (\alpha, \pi)$, so β is an elementary circuit of length $l(\beta) \geqq s+1$ with $l(\beta)$ coprime to s. We consider two cases. First assume that $j-i \leqq s-2$. Since $(l(\beta), s) = 1$, in D^s there exists an elementary circuit connecting the vertices of β. However, β contains $j-i+1 < s$ vertices of $(D \sim \pi)^s$, so the digraph D^* defined by (4.5) is not an elementary circuit. It follows from Lemma 4.7 that

$\gamma(D) \leqq n-1+s(n-3)$. Now assume that $j-i \geqq s-1$, so s consecutive vertices of β are vertices of $D \sim \pi$. None of $y_i, \cdots, y_j$ lies on a circuit of D of length s since D has a unique circuit of length s containing ρ and $\mu(D) \geqq 2$. Since $(l(\beta), s) = 1$, in D^s there exists an elementary circuit containing $y_i, \cdots, y_j$ and the internal vertices of π. This elementary circuit ζ passes through every component of $(D \sim \pi)^s$, but does not contain any loop vertices of D^s. For some $p \in \{1, \cdots, k-1\}$ and $q \in \{i, \cdots, j\}$, there exists an arc $\{x_p, y_q\}$ of ζ. Let A be the connected component of $(D \sim \pi)^s$ which contains y_q, and let $a \in A$ be a loop vertex. Since $s \geqq 3$ and ζ contains no loop vertices, there exists in D^s a path from a to x_p which avoids at least two loop vertices. If some arc with initial vertex x_p has terminal vertex in $X - A$, then the digraph D^* defined by (4.5) is not an elementary circuit, and by Lemma 4.7, $\gamma(D) \leqq n-1+s(n-3)$. Thus we may assume that all arcs of D^s with initial vertex x_p have terminal vertex in A. Now $s \geqq 3$ and the fact that ζ contains no loop vertices imply that there exists a path in D^s from a to any vertex other than x_p which avoids x_p and at least one loop vertex. So $\max\{d_{D^s}(a, x): x \in X\} \leqq n-3$, and by Lemma 4.6 we have $\gamma(D) \leqq n-1+s(n-3)$. The proof of the theorem when $s \geqq 3$ is now complete.

Suppose now $s = 2$; we wish to show that $\gamma(D) \leqq 3n-7$. Let $2, p_1, \cdots, p_t, q$ be the distinct lengths of the elementary circuits of D, where $2 | p_i$ for $i = 1, \cdots, t$, and q is odd. Clearly $F(2, p_1, \cdots, p_t, q) = F(2, q) = q-2$. By (4.6) of Lemma 4.8, $\gamma(D) \leqq F(2, p_1, \cdots, p_t, q) + r(D) + 1$, or

$$\gamma(D) \leqq \max\{q-1+r_{x,y}: x, y \in X\}. \tag{4.10}$$

Let σ be the elementary circuit of D of length 2 and let β be the elementary circuit of D of length q. If one of x, y is a vertex of σ and the other is a vertex of β, then $r_{x,y} \leqq n-1$. In this case $q-1+r_{x,y} \leqq n+q-2 \leqq 2n-4 < 3n-7$ since $n \geqq 4$. If one of x, y is a vertex of σ and the other is not a vertex of β, then a shortest path from x to y which contains a vertex of β need not use an internal vertex of π. Hence $r_{x,y} \leqq 2(n-2)$. Also, since $\mu(D) \geqq 2$ and D is special, σ and β have no vertices in common, so $q \leqq n-2$. Thus $q-1+r_{x,y} \leqq 3n-7$. Now if one of x, y is a vertex of β and the other is not a vertex of σ, a similar argument shows again that $q-1+r_{x,y} \leqq 3n-7$. Finally, suppose neither x nor y is a vertex of either σ or β. There is a path α from x to a vertex v of σ of length at most $n-3$. If α contains a vertex of β, then since there is a path from v to y of length at most $n-2$, we have $r_{x,y} \leqq 2n-5$ and $q-1+r_{x,y} \leqq 3n-8$. If α does not contain a vertex of β, then $l(\alpha) \leqq n-q-2$. Now there is a path from v to a vertex u of β of length at most $n-3$, and a path from u to y of length at most $n-4$. Thus $r_{x,y} \leqq 3n-q-9$, and $q-1+r_{x,y} \leqq 3n-10$. It now follows from (4.10) that $\gamma(D) \leqq 3n-7$. This completes the proof of the theorem. □

5. Gaps in the exponent set. In this section we apply Theorem 4.9 to show that there exist gaps in the exponent set of primitive, minimally strong digraphs on n vertices (primitive, nearly reducible $n \times n$ matrices). Our methods will be similar to those of Dulmage and Mendelsohn [6]. For simplicity, we shall state and prove the results in terms of digraphs. Our first result is Theorem 4.2 of [4], now an easy corollary of Theorem 4.9.

COROLLARY 5.1. *Let D be a primitive, minimally strong digraph on n vertices. Then*

$$\gamma(D) \leqq n^2 - 4n + 6, \tag{5.1}$$

with equality if and only if D is isomorphic to the digraph $D_{n-2,n}$.

Proof. Suppose that D has an elementary circuit of length n. Then by (2.3) of Lemma 2.1, D contains no other arcs, and hence is not primitive. This contradiction shows that D has no elementary circuit of length n.

If $s = s(D) = n-1$, then since D is primitive, D also contains an elementary circuit of length n, a contradiction. So $s \leqq n-2$, and $n + s(n-3) \leqq n + (n-2)(n-3)$, with equality exactly when $s = n-2$. Thus by Theorem 4.9 the inequality (5.1) holds; in addition, equality holds in (5.1) if and only if $D \cong D_{n-2,n}$. □

COROLLARY 5.2. *Let n be an integer at least six. Then there exists no primitive, minimally strong digraph D on n vertices such that either*

$$n^2-5n+9 < \gamma(D) < n^2-4n+6, \quad \text{or} \quad n^2-6n+12 < \gamma(D) < n^2-5n+9.$$

Up to isomorphism, there exist either zero or one primitive, minimally strong digraphs on n vertices with exponent n^2-5n+9, *according to whether n is odd or even. Furthermore, there exist either one or two nonisomorphic primitive, minimally strong digraphs on n vertices with exponent* $n^2-6n+12$, *according to whether* $n-1$ *is or is not a multiple of three.*

Proof. Since the proof is straightforward, we leave it to the reader to check many of the details. Let D be a primitive, minimally strong digraph on n vertices with

$$n^2-6n+12 \leqq \gamma(D) < n^2-4n+6. \tag{5.2}$$

If $s = s(D) \leqq n-4$, then by Theorem 4.9, $\gamma(D) \leqq n^2-6n+12$, with equality if and only if $D \cong D_{n-4,n}$. However, $D_{n-4,n}$ is primitive exactly when $3 \nmid n-1$.

So we assume that $s \geqq n-3$. We saw in the proof of Corollary 5.1 that $s \leqq n-2$. However, if $s = n-2$, then we saw that $D \cong D_{n-2,n}$ and $\gamma(D) = n^2-4n+6$. Thus $s \leqq n-3$, and henceforth we shall assume that $s = n-3$. If D contains a circuit of length $n-1$, then $D \cong D_{n-3,n}$. The digraph $D_{n-3,n}$ is primitive exactly when $2|n$, and it follows from Theorem 4.9 that in this case $\gamma(D) = n^2-5n+9$.

We are left only to consider primitive, minimally strong digraphs on n vertices all of whose elementary circuits have length $n-3$ or $n-2$. First consider the digraph H_n of Fig. 7. This digraph has $r(H_n) = r_{n-4,n-2} = n$, and since the vertices $n-4$, $n-2$ have the unique path property, it follows from Lemma 4.8 that $\gamma(H_n) = n^2-6n+12$.

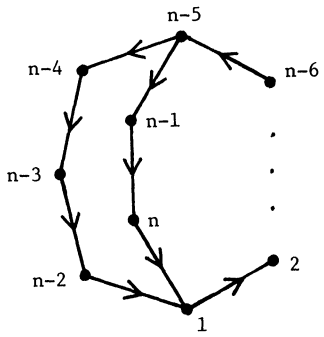


FIG. 7. *The digraph* H_n.

All other primitive, minimally strong digraphs on n vertices which have all their elementary circuits of length $n-3$ or $n-2$ are shown in Fig. 8. If D is isomorphic to one of the digraphs (i), (ii), or (iii), then $r(D) = r_{n-3,n-2} = n-1$. Since in each case the vertices $n-3$, $n-2$ have the unique path property, it follows from Lemma 4.8 that $\gamma(D) = n^2-6n+11$. If D is isomorphic to one of the digraphs (iv) or (v), then

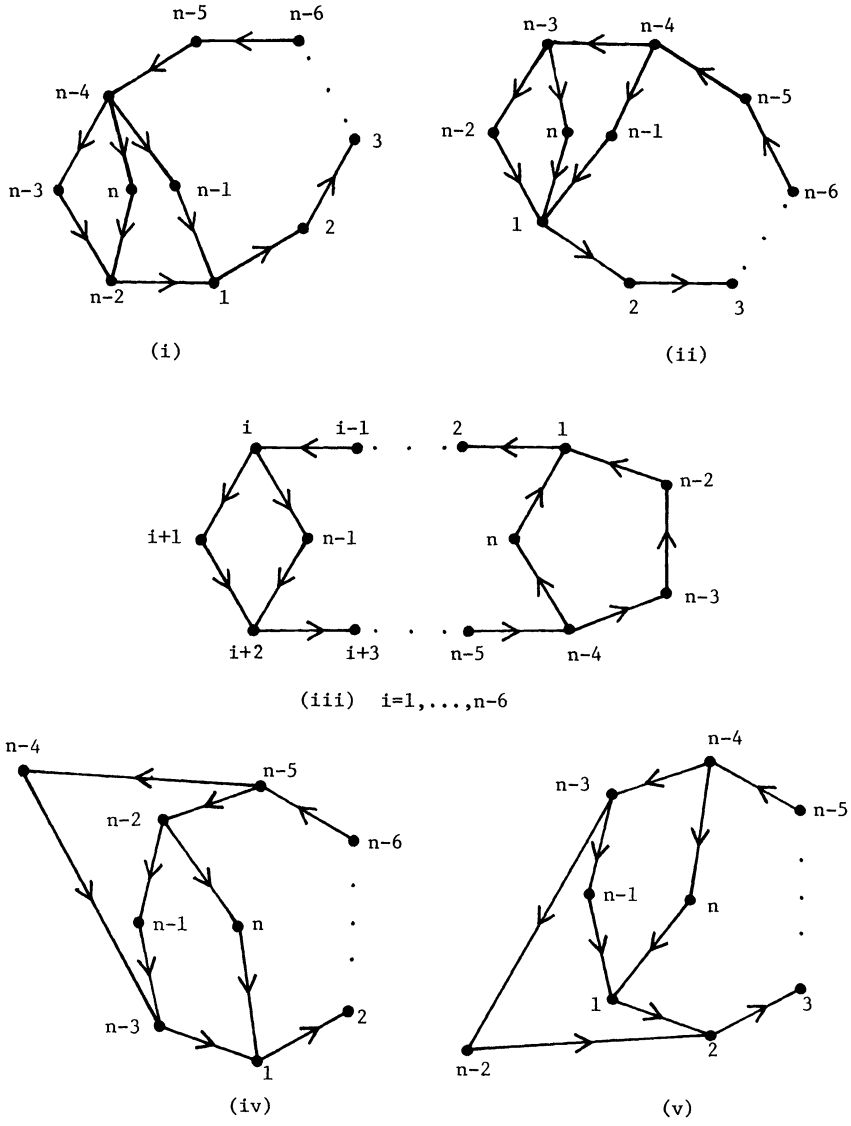


FIG. 8

$r(D) = r_{n-1,n-1} = n-2$. Again the vertices $n-1$, $n-1$ have the unique path property, and by Lemma 4.8 we see that $\gamma(D) = n^2 - 6n + 10$. □

We summarize the results of this section by listing all primitive, minimally strong digraphs on $n \geqq 6$ vertices with exponent at least $n^2 - 6n + 12$.

Digraph		Exponent
$D_{n-2,n}$		n^2-4n+6
$D_{n-3,n}$	(for n even)	n^2-5n+9
H_n		$n^2-6n+12$
$D_{n-4,n}$	(for $3 \nmid n-1$)	$n^2-6n+12$

6. Concluding remarks. We conclude this paper by discussing some open problems concerning the exponent of a primitive, nearly reducible matrix, and concerning the exponent of a primitive matrix in general. In § 3 we saw that there exists an $n \times n$ primitive, nearly reducible matrix with exponent k whenever $6 \leqq k \leqq n+1$. On the other hand, in § 5 we saw that for $n \geqq 5$, hence for all positive n, there does not exist an $n \times n$ primitive, nearly reducible matrix with exponent $n^2-6n+13$. These results lead to the following problems.

Problem 6.1. For $n \geqq 5$, what is the least integer $e(n) \geqq 6$ such that no $n \times n$ primitive, nearly reducible matrix has exponent $e(n)$?

That for $n \geqq 5$ there exist gaps in the exponent set of $n \times n$ primitive matrices was shown by Dulmage and Mendelsohn [6]. Hence we may also ask the following.

Problem 6.2. What is the least positive integer $e_1(n)$ such that no $n \times n$ primitive matrix has exponent $e_1(n)$?

Some of the difficulty in the proof of Theorem 4.9 arose in the effort to characterize equality in (4.8). Primitive matrices for which equality holds in (1.1) have not been characterized. Dulmage and Mendelsohn [6] have noted that if $(s, n) = 1$, then the digraph $K_{s,n}$ of Fig. 9 is primitive with $s(K_{s,n}) = s$ and $\gamma(K_{s,n}) = n + s(n-2)$.

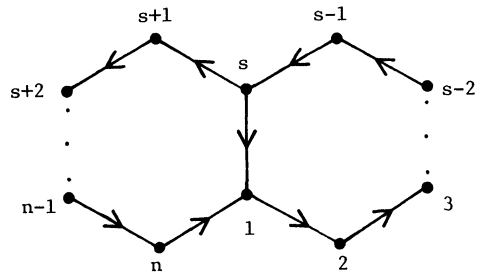


FIG. 9. *The digraph* $K_{s,n}$.

Problem 6.3. Characterize the primitive digraphs on n vertices with shortest circuit length s and exponent $n + s(n-2)$.

When $(s, n) = 1$, it is not always the case that $K_{s,n}$ is the only primitive digraph on n vertices with shortest circuit length s and exponent $n + s(n-2)$. For example, the digraph of Fig. 10 is primitive on 10 vertices, has shortest circuit of length 3, and

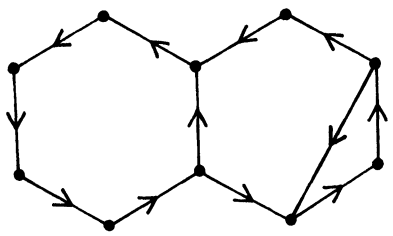


FIG. 10

has exponent $34 = 10 + 3(10-2)$. Notice that the elementary circuits of this digraph have lengths 3, 6, and 10, and $F(3, 6, 10) = F(3, 10) = 17$. This example is not unique; using the same technique, it may be possible to construct a primitive digraph on n vertices with elementary circuits of length $s = p_1 < p_2 < \cdots < p_t = n$ and exponent

$n+s(n-2)$. In this case $F(p_1, \cdots, p_t) = F(p_1, p_t)$. One may check that as long as $F(p_1, \cdots, p_t) = F(s, n)$, $p_i | sn$ for $i = 1, \cdots, t$, and $s + p_{t-1} < n$, then a primitive digraph on n vertices with elementary circuits of lengths $p_1, \cdots, p_t$ may be constructed in a manner similar to the digraph of Fig. 10. However, this sort of construction always yields a digraph which is an elementary circuit on n vertices with some chords. Based on this, we may ask a simpler question than Problem 6.3.

Problem 6.4. If D is a primitive digraph on n vertices with $s(D) = s \geqq 2$ and $\gamma(D) = n + s(n-2)$, does D contain an elementary circuit of length n?

In Problem 6.4 we required $s \geqq 2$. Let D be a strong digraph on n vertices which contains a loop, so D is primitive and $\gamma(D) \leqq 2n-2$. If $\gamma(D) = 2n-2$, then D need not contain an elementary circuit of length n. In particular, the digraphs $B_{i,n}$ of Fig. 11, $i = 1, \cdots, n-1$, may have exponent $2n-2$, but only $B_{1,n}$ has an elementary circuit of length n.

Problem 6.5. Let D be a strong digraph on n vertices which contains a loop, and suppose D is not isomorphic to $B_{i,n}$ for $i = 2, \cdots, n-1$. If $\gamma(D) = 2n-2$, does D contain an elementary circuit of length n?

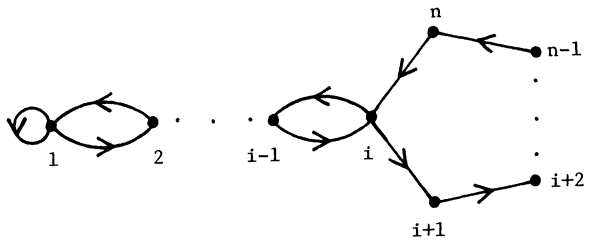


FIG. 11. *The digraph* $B_{i,n}$.

Note added in proof. A system of gaps in the exponent set of primitive matrices, Illinois J. Math., 25 (1981), pp. 87–98, by M. Lewin and Y. Vitek, has recently come to the author's attention. In this paper it is conjectured that $e_1(n) \geqq (n^2-2n+4)/2$.

REFERENCES

[1] C. BERGE, *Graphs and Hypergraphs*, North-Holland, Amsterdam, 1973.
[2] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.
[3] R. A. BRUALDI AND M. B. HEDRICK, *A unified treatment of nearly reducible and nearly decomposable matrices*, Linear Algebra Appl., 24 (1979), pp. 51–73.
[4] R. A. BRUALDI AND J. A. ROSS, *On the exponent of a primitive, nearly reducible matrix*, Math. Oper. Res., 5 (1980), pp. 229–241.
[5] E. V. DENARDO, *Periods of connected networks and powers of nonnegative matrices*, Math. Oper. Res., 2 (1977), pp. 20–24.
[6] A. L. DULMAGE AND N. S. MENDELSOHN, *Gaps in the exponent set of primitive matrices*, Illinois J. Math., 8 (1964), pp. 642–656.
[7] B. R. HEAP AND M. S. LYNN, *The index of primitivity of a nonnegative matrix*, Numer. Math., 6 (1964), pp. 120–141.
[8] J. C. HOLLADAY AND R. S. VARGA, *On powers of nonnegative matrices*, Proc. Amer. Math. Soc., 9 (1958), pp. 631–634.
[9] J. G. KEMENY AND J. L. SNELL, *Finite Markov Chains*, Van Nostrand, Princeton, NJ, 1960.
[10] M. LEWIN, *On exponents of primitive matrices*, Numer. Math., 18 (1971), pp. 154–161.
[11] ———, *Bounds for exponents of doubly stochastic primitive matrices*, Math. Z., 137 (1974), pp. 21–30.
[12] J. A. ROSS AND C. LUCCHESI, *Superfluous paths in strong digraphs*, submitted.
[13] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.
[14] H. WIELANDT, *Unzerlegbare, nicht negativen matrizen*, Math. Z., 52 (1950), pp. 642–645.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0001 $01.00/0

REARRANGEABLE NETWORKS WITH LIMITED DEPTH*

NICHOLAS PIPPENGER† AND ANDREW C.-C. YAO‡

Abstract. Rearrangeable networks are switching systems capable of establishing simultaneous independent communication paths in accordance with any one-to-one correspondence between their n inputs and n outputs. Classical results show that $\Omega(n \log n)$ switches are necessary and that $O(n \log n)$ switches are sufficient for such networks. We are interested in the minimum possible number of switches in rearrangeable networks in which the depth (the length of the longest path from an input to an output) is at most k, where k is fixed as n increases. We show that $\Omega(n^{1+1/k})$ switches are necessary and that $O(n^{1+1/k}(\log n)^{1/k})$ switches are sufficient for such networks.

1. Introduction. An *(m, n)-network* $G = (V, E, A, B)$ comprises an acyclic directed graph with *vertices* V and *edges* E, a set of m distinguished vertices A called *inputs* and a set of n other distinguished vertices B called *outputs*.

A *request* is an ordered pair (a, b) comprising an input a and an output b. A *route* is a directed path from an input to an output. A route *satisfies* a request (a, b) if it is from a to b.

An *l-assignment* is a set of l requests, no two of which have an input or output in common. An *l-state* is a set of l routes, no two of which have a vertex in common. An l-state *satisfies* an l-assignment if it contains a route satisfying each request in the assignment.

An *n-connector* (also known as a *rearrangeable* n-network) is an (n, n)-network that has an n-state satisfying each of the $n!$ n-assignments. The *size* of a network is the number of edges in it. The *depth* of a network is the maximum number of edges in any route in it.

Let $f(n)$ denote the minimum possible size of an n-connector. An information-theoretic argument (due to C. E. Shannon) shows that $f(n) = \Omega(n \log n)$ (see Pippenger [4]; $\Omega(\cdots)$ means "some function bounded below by a strictly positive constant times $\cdots$"). A classical construction (due to D. Slepian, A. M. Duguid and J. LeCorre) shows that $f(n) = O(n \log n)$ (see Pippenger [3]; $O(\cdots)$ means "some function bounded above in absolute value by a constant times $\cdots$").

Let $f_k(n)$ denote the minimum possible size of an n-connector having depth at most k. We shall be interested in the behavior of $f_k(n)$ as n grows while k remains fixed. The case $k = 1$ is trivial: $f_1(n) = n^2$. For $k = 2$, a probabilistic argument (used by de Bruijn, Erdös and Spencer [1] to solve a problem of van Lint [2]) shows that $f_2(n) = O(n^{3/2}(\log n)^{1/2})$. For odd $k \geqq 3$, the classical construction referred to above shows that $f_k(n) = O(n^{1+2/(k+1)})$.

In § 2 we shall show (by adapting an argument due to Pippenger and Valiant [5]) that $f_k(n) = \Omega(n^{1+1/k})$. In §§ 3 and 4 we shall show (by a probabilistic argument) that $f_k(n) = O(n^{1+1/k}(\log n)^{1/k})$.

2. Lower bound. An *n-tree* is a $(1, n)$-network with inputs $A = \{a\}$, outputs $B = \{b_1, \cdots, b_n\}$ and, for $1 \leqq j \leqq n$, a unique route R_j satisfying the request (a, b_j).

If T is an n-tree, let

$$\Delta(T) = \sum_{1 \leqq j \leqq n} \sum_{v \in R_j} d_v,$$

* Received by the editors June 6, 1981.

† Computer Science Department, IBM Research Laboratory, San Jose, California 95193.

‡ Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720.

where d_v denotes the number of edges directed out of the vertex v.

PROPOSITION 2.1. *If T is an n-tree of depth at most $k \geqq 1$, then*

$$\Delta(T) \geqq kn^{1+1/k}.$$

Proof. The proof is by induction on k. The case $k = 1$, $\Delta(T) = n^2$ is trivial. If $k \geqq 2$, let d be the number of edges directed out of the input and let $T_1, \cdots, T_d$ (with $n_1, \cdots, n_d$ outputs, respectively) be the subtrees into which these edges are directed.

We have

$$\Delta(T) = dn + \sum_{1 \leqq h \leqq d} \Delta(T_h) \geqq dn + \sum_{1 \leqq h \leqq d} (k-1) n_h^{1+1/(k-1)},$$

by inductive hypothesis. Since $n_1 + \cdots + n_d = n$ and $(k-1)\theta^{1+1/(k-1)}$ is convex in θ, we have

$$\sum_{1 \leqq h \leqq d} (k-1) n_h^{1+1/(k-1)} \geqq d(k-1)\left(\frac{n}{d}\right)^{1+1/(k-1)}.$$

Straightforward calculus shows that

$$dn + d(k-1)\left(\frac{n}{d}\right)^{1+1/(k-1)} \geqq kn^{1+1/k},$$

which completes the induction. □

An *n-shifter* is an (n, n)-network with inputs $A = \{a_1, \cdots, a_n\}$, outputs $B = \{b_1, \cdots, b_n\}$ and, for $1 \leqq j \leqq n$, a state satisfying the assignment $\{(a_1, b_{j+1}), \cdots, (a_n, b_{j+n})\}$ (addition is modulo n).

THEOREM 2.1. *Any n-shifter of depth at most k has size at least $kn^{1+1/k}$.*

Proof. Let $G = (V, E, A, B)$ be an n-shifter of depth at most k. Let $R_{i,j}$ be the route from input a_i to output b_{j+i} in the state that satisfies the assignment $\{(a_1, b_{j+1}), \cdots, (a_n, b_{j+n})\}$. By identifying common initial segments, the routes $R_{i,1}, \cdots, R_{i,n}$ can be assembled into an n-tree T_i of depth at most k, for which

$$\Delta(T_i) \geqq kn^{1+1/k},$$

by Proposition 2.1. For $1 \leqq i \leqq n$, $1 \leqq j \leqq n$ and $e \in E$, let $\mu(i, j, e)$ be 1 if the edge e is directed out of a vertex on $R_{i,j}$ and 0 otherwise. For any i and j, we have

$$\sum_{e \in E} \mu(i, j, e) \geqq \sum_{v \in R_{i,j}} d_v,$$

and by summing over j, we have

$$\sum_{1 \leqq j \leqq n} \sum_{e \in E} \mu(i, j, e) \geqq \sum_{1 \leqq j \leqq n} \sum_{v \in R_{i,j}} d_v \geqq \Delta(T_i) \geqq kn^{1+1/k}.$$

Thus,

$$\sum_{1 \leqq i \leqq n} \sum_{1 \leqq j \leqq n} \sum_{e \in E} \mu(i, j, e) \geqq kn^{2+1/k}. \tag{2.1}$$

On the other hand, since the routes $R_{1,j}, \cdots, R_{n,j}$ have no vertex in common, an edge e can be directed out of a vertex on at most one of them. Thus, for any j and e, we have

$$\sum_{1 \leqq i \leqq n} \mu(i, j, e) \leqq 1.$$

By summing over j we have

$$\sum_{1\leqq j\leqq n}\sum_{1\leqq i\leqq n}\mu(i,j,e)\leqq n,$$

and by summing over e we have

$$\sum_{e\in E}\sum_{1\leqq j\leqq n}\sum_{1\leqq i\leqq n}\mu(i,j,e)\leqq n\,\#(E)$$

($\#(\cdots)$ means "the cardinality of $\cdots$"). Comparing this with (2.1) gives

$$\#(E)\geqq kn^{1+1/k},$$

as claimed. □

COROLLARY 2.1. *Any n-connector of depth at most k has size at least $kn^{1+1/k}$.*

Proof. An n-connector is an n-shifter. □

3. Couplers. A set $\mathscr{X}=\{X_1,\cdots,X_r\}$ is an *x-packing* of a set A if $X_1,\cdots,X_r$ are mutually disjoint x-element subsets of A. An x-packing $\mathscr{X}$ of A is *tight* if $\#(\bigcup\mathscr{X})\geqq \#(A)/16$ (or, equivalently, $\#(\mathscr{X})\geqq\#(A)/16x$).

If G is a network and X a set of inputs of G, let $G(X)$ denote the set of outputs of G reachable through routes from inputs in X.

An (l,l)-networks $G=(V,E,A,B)$ is an *(l,x,y)-coupler* if, for every tight x-packing $\mathscr{X}=\{X_1,\cdots,X_r\}$ of A, there exists a tight y-packing $\mathscr{Y}=\{Y_1,\cdots,Y_s\}$ of B such that, for every $1\leqq j\leqq s$, there exists $1\leqq i\leqq r$ such that $Y_j\subseteq G(X_i)$.

If G is an (l,m)-network and H is an (m,n)-network, let $G\circ H$ denote an (l,n)-network obtained by identifying the outputs of G with the inputs of H in any one-to-one fashion (to become vertices that are neither inputs nor outputs of $G\circ H$).

LEMMA 3.1. *If G is an (l,x,y)-coupler and H is an (l,y,z)-coupler, then $G\circ H$ is an (l,x,z)-coupler.*

Proof. The proof is immediate. □

An (m,m)-network G is a *strong (m,x,y)-coupler* if, for every $m/2\leqq l\leqq m$, each (l,l)-network obtained from G by deleting $m-l$ vertex-disjoint routes (together with all edges incident with vertices on these routes) is an (l,x,y)-coupler.

LEMMA 3.2. *If G is a strong (m,x,y)-coupler and H is a strong (m,y,z)-coupler, then $G\circ H$ is a strong (m,x,z)-coupler.*

Proof. The proof follows from Lemma 3.1. □

LEMMA 3.3. *Let $\mathbf{X}$ denote the number of successes among n trials that succeed independently with probability p. Then*

$$\mathscr{P}(\mathbf{X}>2np)\leqq\left(\frac{e}{4}\right)^{np} \tag{3.1}$$

and

$$\mathscr{P}\left(\mathbf{X}<\frac{np}{2}\right)\leqq\left(\frac{2}{e}\right)^{np/2} \tag{3.2}$$

($\mathscr{P}(\cdots)$ *means "the probability of* $\cdots$").

Proof. For (3.1), we may assume $p<\frac{1}{2}$, for otherwise $\mathscr{P}(\mathbf{X}>2np)=0$. If

$$\mathbf{Y}=\begin{cases}0 & \text{if } \mathbf{X}\leqq 2np,\\ 1 & \text{if } \mathbf{X}>2np,\end{cases}$$

then $\mathscr{P}(\mathbf{X}>2np)=\mathscr{E}(\mathbf{Y})$. If $\mathbf{Z}=T^{\mathbf{X}-2np}$ (where $T>1$ is a parameter to be chosen later), then $\mathbf{Y}\leqq\mathbf{Z}$ and so $\mathscr{E}(\mathbf{Y})\leqq\mathscr{E}(\mathbf{Z})$. Thus it will suffice to estimate $\mathscr{E}(\mathbf{Z})$.

Since $\mathbf{X}$ is the sum of n independent random variables that assume the value 1 with probability p and the value 0 with probability $1-p$, $T^{\mathbf{X}}$ is the product of n independent random variables that have expected value $pT+1-p$. Thus,

$$\mathscr{E}(\mathbf{Z}) = (pT+1-p)^n T^{-2np}.$$

Choosing $T = 2(1-p)/(1-2p)$ and using the inequality $1+\theta \leqq e^{\theta}$ yields (3.1). A similar argument yields (3.2). □

PROPOSITION 3.1. If

$$512x \ln m \leqq y \leqq \frac{m}{16},$$

then there exists a strong (m, x, y)*-coupler of depth* 1 *and size at most* $32my/x$.

Proof. Let

$$p = \frac{16y}{x},$$

and let $\mathbf{G} = (V, \mathbf{E}, A, B)$ be the random (m, m)-network $\mathbf{K}_{m,m}(p)$ (an (m, m)-network of depth 1 in which each of the m^2 potential edges is independently present with probability p). We expect $m^2 p = 16my/x$ edges in $\mathbf{E}$, so

$$\mathscr{P}\left(\#(\mathbf{E}) > \frac{32my}{x}\right) \leqq \left(\frac{e}{4}\right)^{16my/x} \leqq \frac{1}{4},$$

by Lemma 3.3. Thus, it will suffice to show that

$$\mathscr{P}(\mathbf{G} \text{ not a strong } (m, x, y)\text{-coupler}) \leqq \frac{1}{4}.$$

There are at most 4^m (l, l)-networks $\mathbf{F}$ obtained from $\mathbf{G}$ by deleting $m-l$ vertex-disjoint routes (together with all edges incident with vertices on these routes). It will thus suffice to show for $m/2 \leqq l \leqq m$ and $\mathbf{F} = \mathbf{K}_{l,l}(p)$ that

$$\mathscr{P}(\mathbf{F} \text{ not an } (l, x, y)\text{-coupler}) \leqq \frac{1}{4^{m+1}}.$$

There are at most $l^l \leqq m^m$ minimal tight x-packings $\mathscr{X} = \{X_1, \cdots, X_r\}$ (where $r = \lceil l/16x \rceil$) of the inputs of $\mathbf{F}$. Thus, it will suffice to show that

$$\mathscr{P}(\mathbf{F} \text{ not an } (l, x, y)\text{-coupler for } \mathscr{X}) \leqq \frac{1}{4^{m+1} m^m}.$$

We shall consider each set X_i in turn. For each set X_i, we shall attempt to construct a set Y_j containing y outputs, disjoint from all previously constructed sets $Y_1, \cdots, Y_{j-1}$ and satisfying $Y_j \subseteq \mathbf{F}(X_i)$. If we show that

$$\mathscr{P}(\text{no } Y_j \text{ for } X_i) \leqq \left(\frac{2}{e}\right)^y,$$

then the probability of fewer than $s = \lceil l/16y \rceil$ successes among r trials will be at most

$$2^r \left[\left(\frac{2}{e}\right)^y\right]^{r-s} \leqq 2^r \left[\left(\frac{2}{e}\right)^y\right]^{r/2} \leqq \left[4\left(\frac{2}{e}\right)^y\right]^{l/32x} \leqq \left[4\left(\frac{2}{e}\right)^y\right]^{m/64x}$$

$$\leqq 4^m \left(\frac{2}{e}\right)^{my/64x} \leqq 4^m e^{-my/256x} \leqq 4^m e^{-2m \ln m} \leqq \frac{1}{4^{m+1} m^m}.$$

To construct Y_j, we shall consider each output of $\mathbf{F}$ that is not in $Y_1 \cup \cdots \cup Y_{j-1}$ in turn (there are at least $l - sy \geqq l/2 \geqq m/4$ such outputs). For each such output, we shall attempt to find an edge joining it to an input in X_i. The probability of finding such an edge is

$$1-(1-p)^x \geqq 1-e^{-px} \geqq \frac{px}{2}$$

(using $1-\theta \leqq e^{-\theta} \leqq 1-\theta/2$ for $0 \leqq \theta \leqq 1$). Thus, we expect at least $(m/4)(px/2) = 2y$ successes and the probability of fewer than y successes is at most $(2/e)^y$, by Lemma 3.3. □

COROLLARY 3.1. *If*

$$512 \ln m \leqq x$$

and

$$x^{k-1} \leqq \frac{m}{16},$$

then there exists a strong (m, x, x^{k-1})*-coupler of depth* $k-2$ *and size at most* $32(k-2)mx$.

Proof. The proof follows from Lemma 3.2 and Proposition 3.1. □

4. Upper bound. An (n, n)-network is an (a, b)*-partial n-connector* if, for every a-assignment P, there exists an $(a-b)$-assignment $Q \subseteq P$ and a state satisfying Q.

If Q and H are (n, n)-networks, let $G \| H$ denote an (n, n)-network obtained by identifying the inputs of G with the inputs of H in any one-to-one fashion (to become the inputs of $G \| H$) and identifying the outputs of G with the outputs of H in any one-to-one fashion (to become the outputs of $G \| H$).

LEMMA 4.1. *If* G *is an* (a, b)*-partial n-connector and* H *is a* (b, c)*-partial n-connector, then* $G \| H$ *is an* (a, c)*-partial n-connector.*

Proof. The proof is immediate. □

LEMMA 4.2. *Let* L *be an* l*-element set and let* $\mathscr{C}$ *be a collection of subsets of* L *such that if* $Y \in \mathscr{C}$ *and* $X \subseteq Y$, *then* $X \in \mathscr{C}$. *If* $\mathscr{C}$ *contains more than*

$$2^{-2x}\binom{l}{2x}$$

$(2x)$*-element subsets of* L, *then it contains a tight* x*-packing of* L.

Proof. Let $\mathbf{Y}$ be a random uniformly distributed $(2x)$-element subset of L, then

$$\mathscr{P}(\mathbf{Y} \in \mathscr{C}) > 2^{-2x}.$$

Let $\mathscr{X}$ be a maximal x-packing contained in $\mathscr{C}$. If $\mathscr{X}$ is not tight, then

$$\#(\bigcup \mathscr{X}) \leqq \frac{l}{16}$$

and

$$\mathscr{P}(\#(\mathbf{Y} \cap \bigcup \mathscr{X}) > x) \leqq 2^{2x} 16^{-x} \leqq 2^{-2x}.$$

Thus, there exists a $(2x)$-element set $Y \in \mathscr{C}$ for which $\#(Y \cap \bigcup \mathscr{X}) \leqq x$. Then $Y - \bigcup \mathscr{X}$ contains an x-element set that can be added to $\mathscr{X}$ to yield a larger x-packing, contradicting the maximality of $\mathscr{X}$. □

PROPOSITION 4.1. *If*

$$512^k (2 \ln 2n)^{k-1} \leqq m \leqq n,$$

then there exists an $(m, m/2)$*-partial* n*-connector of depth* k *and size at most* $64(k-1)\cdot n(2m \ln 2n)^{1/k}$.

Proof. Set

$$x = \lceil (2m \ln 2n)^{1/k} \rceil.$$

Then $512 \ln m \leqq x$ and $x^{k-1} \leqq m/16$, and by Corollary 3.1, there exists a strong (m, x, x^{k-1})-coupler G of depth $k-2$ and size $32(k-2)mx$.

Let

$$q = \frac{8x}{m},$$

and let $\mathbf{H} = (V, \mathbf{E}, A, B)$ be the random (n, n)-network $\mathbf{K}_{n,m}(q) \circ G \circ \mathbf{K}_{m,n}(q)$. We expect $2nmq = 16nx$ edges in $\mathbf{K}_{n,m}(q)$ and $\mathbf{K}_{m,n}(q)$ together, so

$$\mathcal{P}(\#(\mathbf{E}) \geqq 32(k-1)nx) \leqq \left(\frac{e}{4}\right)^{16nx} \leqq \frac{1}{4},$$

by Lemma 3.3. It will thus suffice to show that

$$\mathcal{P}(\mathbf{H} \text{ not an } (m, m/2)\text{-partial } n\text{-connector}) \leqq \tfrac{1}{4}.$$

There are at most $n^{2m}/4$ m-assignments P. Thus, it will suffice to show that

$$\mathcal{P}(\mathbf{H} \text{ not an } (m, m/2)\text{-partial } n\text{-connector for } P) \leqq n^{-2m}.$$

We shall consider each of the m requests in P in turn. For each request (a, b), we shall attempt to construct a route, vertex-disjoint from all previously constructed routes and satisfying the request (a, b). If we show that

$$\mathcal{P}(\text{no route for } (a, b)) \leqq \frac{1}{4n^4},$$

then the probability of fewer than $m/2$ successes among m trials will be at most

$$2^m \left(\frac{1}{4n^4}\right)^{m/2} = n^{-2m}.$$

The probability that there is no route for (a, b) is the probability that there is no route in the random network $\mathbf{I} = \mathbf{K}_{1,l}(q) \circ F \circ \mathbf{K}_{l,1}(q)$, where $m/2 \leqq l \leqq m$ and F is an (l, x, x^{k-1})-coupler. Let ξ denote the random number of outputs of $\mathbf{K}_{1,l}(q)$ reachable through routes from the input of $\mathbf{K}_{1,l}(q)$, let η denote the random number of outputs of $\mathbf{K}_{1,l}(q) \circ F$ reachable through routes from the input of $\mathbf{K}_{1,l}(q) \circ F$ and let ζ denote the random number of inputs of $\mathbf{K}_{l,1}(q)$ from which the output of $\mathbf{K}_{l,1}(q)$ is reachable. Then

$$\begin{aligned}\mathcal{P}(\text{no route}) &\leqq \mathcal{P}(\text{no route} \mid \eta \geqq x^{k-1}, \zeta \geqq 2x) + \mathcal{P}(\eta < x^{k-1}) + \mathcal{P}(\zeta < 2x) \\ &\leqq \mathcal{P}(\text{no route} \mid \eta \geqq x^{k-1}, \zeta \geqq 2x) + \mathcal{P}(\eta < x^{k-1} \mid \xi \geqq 2x) + \mathcal{P}(\xi < 2x) \\ &\quad + \mathcal{P}(\zeta < 2x).\end{aligned}$$

The random variables ξ and ζ have expected value $lq \geqq mq/2 = 4x$, so

$$\mathcal{P}(\xi < 2x) = \mathcal{P}(\zeta < 2x) \leqq \left(\frac{2}{e}\right)^{2x} \leqq \frac{1}{16n^4},$$

by Lemma 3.3. Furthermore,

$$\mathscr{P}(\text{no route} \mid \eta \geqq x^{k-1}, \zeta \geqq 2x) \leqq \binom{l-x^{k-1}}{2x} \Big/ \binom{l}{2x}$$

$$\leqq e^{-2x^k/l} \leqq e^{-2x^k/m} \leqq e^{-4 \ln 2n} \leqq \frac{1}{16n^4}.$$

Thus, it will suffice to show that $\mathscr{P}(\eta < x^{k-1} \mid \xi \geqq 2x) \leqq 1/16n^4$.

Let $\mathscr{C}$ be the collection of subsets X of the inputs of F for which $\#(F(X)) < x^{k-1}$. Since F is an (l, x, x^{k-1})-coupler, $\mathscr{C}$ contains no tight x-packing and thus, by Lemma 4.2, $\mathscr{C}$ contains at most

$$2^{-2x}\binom{l}{2x}$$

$(2x)$-element subsets. Thus,

$$\mathscr{P}(\eta < x^{k-1} \mid \xi \geqq 2x) \leqq 2^{-2x} \leqq \frac{1}{16n^4},$$

as was to be shown. □

COROLLARY 4.1. *If*

$$b = 512^k (2 \ln 2n)^{k-1},$$

then there exists an (n, b)*-partial* n*-connector of depth* k *and size at most*

$$256k(k-1)n(2n \ln 2n)^{1/k}.$$

Proof. The result follows from Lemma 4.1, Proposition 4.1 and

$$\sum_{1 \leqq j < \infty} 2^{-j/k} = \frac{1}{(1-2^{1/k})} \leqq 4k$$

(using $e^{-\theta} \leqq 1 - \theta/2$ for $0 \leqq \theta \leqq 1$). □

LEMMA 4.3. *There exists an* $(a, 0)$*-partial* n*-connector of depth* 2 *and size* $2an$.

Proof. Consider $K_{n,a} \circ K_{a,n}$. □

THEOREM 4.1. *There exists an* n*-connector of depth* k *and size at most*

$$256k(k-1)n(2n \ln 2n)^{1/k} + 2(512)^k n(2 \ln n)^{k-1}.$$

Proof. The result follows from Lemma 4.1, Corollary 4.1 and Lemma 4.3; an $(n, 0)$-partial n-connector is an n-connector. □

REFERENCES

[1] N. G. DE BRUIJN, P. ERDÖS AND J. SPENCER, *Solution* 350, Nieuw Archief voor Wiskunde, 22 (1974), p. 94–109.

[2] J. H. VAN LINT, *Problem* 350, Nieuw Archief voor Wiskunde, 21 (1973), p. 179.

[3] N. PIPPENGER, *On rearrangeable and non-blocking switching networks*, J. Comput. System Sci., 17, (1978), pp. 145–162.

[4] ———, *A new lower bound for the number of switches in rearrangeable networks*, this Journal, 1 (1980), pp. 164–167.

[5] N. PIPPENGER AND L. G. VALIANT, *Shifting graphs and their applications*, J. Assoc. Comput. Mach., 23 (1976), pp. 423–432.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0002 $01.00/0

GOSSIPING WITHOUT DUPLICATE TRANSMISSIONS*

DOUGLAS B. WEST†

Abstract. n people have distinct bits of information, which they communicate via telephone calls in which they transmit everything they know. We require that no one ever hear the same piece of information twice. In the case 4 divides n, $n \geqq 8$, we provide a construction that transmits all information using only $9n/4-6$ calls. Previous constructions used $\frac{1}{2}n \log n$ calls.

The original gossip problem asks for the minimum number of calls permitting a complete passage of information from each person to every other in some group. The answer of $2n-4$ for $n \geqq 4$ has been demonstrated in numerous ways, e.g., [1], and the optimal solutions have been characterized [2], [3]. In [5], we added an additional requirement, that no one hear his own original piece of information in the course of the calling scheme. This is impossible to achieve if n if odd, but if n is even, $2n-4$ calls still suffice, and [5] characterized these solutions.

Next we can prohibit anyone hearing any given piece of information more than once. This implies that no-one hears his own information. If $n \equiv 2 \bmod 4$, then whether it is ever possible to transmit all information under this restriction remains an open question. ($n = 6$ or 10 can be shown impossible without much difficulty.) For 4 divides n, H. W. Lenstra et al. [4] provided an inductive construction that succeeds. If $n/4 \equiv -k \bmod 4$, they divide the people into three groups of size $n/4+k$, $n/4+k$ and $n/2-2k$, each divisible by 4. Forming $n/4$ mini-groups of four people with two from one group and one from the other two, they perform three calls on each. This is done so that in each of the three large groups, all n pieces of information are known by exactly one person. Then they perform induction. If $f(n)$ is the number of calls used, this gives $f(n) = 3n/4 + 2f(n/4+k) + f(n/2-2k)$. This is satisfied by $f(n) \approx \frac{1}{2}n \log n$. (That is exactly the solution if n is a power of 2.)

In this note, we provide an explicit construction for $n \geqq 8$, using only $9n/4-6$ calls. It would be nice to show this is optimal. The best current lower bound is $2n-3$ for $n > 8$, as remarked in [6].

The construction. We begin by dividing the people into $n/4$ groups of 4. In each group, we perform four calls in a square so that each knows all four tidbits from his group. Label the points x_{ij} for $1 \leqq i \leqq n/4$, $1 \leqq j \leqq 4$.

Arrange the squares around a circle, with two points on the inner ring and two on the outer, as in Fig. 1a. We will leave the outer points as they are, knowing 4 pieces of information, until the end. The points on the inner ring will accumulate $n-4$ pieces in such a way that they can then be matched to the outer points.

Label the points in the ith square x_{i1}, x_{i2}, x_{i3}, x_{i4}, so that x_{i1} and x_{i2} are on the inner circle. $x_{1,1}$ and $x_{n/4,1}$ will be special points. We perform in order the calls $(x_{1,2}, x_{2,1})$, $(x_{1,2}, x_{3,1})$, $\cdots$, $(x_{1,2}, x_{n/4-1,1})$ and, also in order, the calls $(x_{n/4,1}, x_{n/4-1,2})$, $(x_{n/4,1}, x_{n/4-2,2})$, $\cdots$, $(x_{n/4,1}, x_{2,2})$. (See Fig. 1b.) In each sequence, four additional bits of information are involved on each call. For $1 < k < n/4$, afterwards $x_{k,1}$ knows all information in $\{x_{ij}: i \leqq k, 1 \leqq j \leqq 4\}$ and $x_{k,2}$ knows all in $\{x_{ij}: i \geqq k, 1 \leqq j \leqq 4\}$, $x_{1,1}$ and $x_{n/4,2}$ still know the four bits they began with, while $x_{1,2}$ knows everything except

* Received by the editors June 8, 1981. This research was supported in part by the National Science Foundation under grant MCS-77-23738 and by the Office of Naval Research under contracts N00014-76-C-0330 and N00014-76-C-0688.

† Mathematics Department, Princeton University, Princeton, New Jersey 08544.

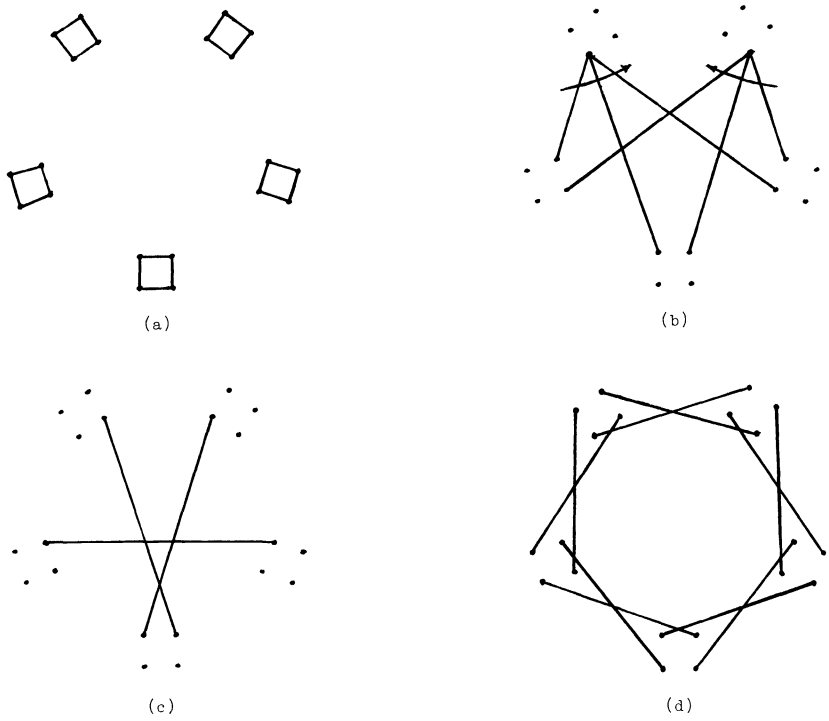


FIG. 1

$\{x_{n/4,j}\}$, and $x_{n/4,1}$ knows everything except $\{x_{1j}\}$. Note that four points $x_{1,2}$, $x_{n/4-1,1}$, $x_{n/4,1}$ and $x_{2,2}$ already know $n-4$ pieces of information.

In the third phase, $x_{k-1,1}$ and $x_{k+1,2}$ call each other, for $2 \leqq k \leqq n/4-1$. (See Fig. 1c.) The former knows the "lowest" $4(k-1)$ pieces of information and the latter the "highest" $4(n/4-k)$ pieces. Together they now know all but $\{x_{kj}: 1 \leqq j \leqq 4\}$.

Finally, the two inside points, knowing all but $\{x_{kj}\}$, are matched with the two outside points, knowing only $\{x_{kj}\}$, for $1 \leqq k \leqq n/4$. This completes the construction.

It is easy to see that no pair of points both knowing any given piece of information ever speak to each other, so there are no duplicate transmissions, and at the end everyone knows everything. Summing up the number of calls used in each of the four stages, we have $n+2(n/4-2)+(n/4-2)+n/2 = 9n/4-6$ total calls.

REFERENCES

[1] B. BAKER AND R. SHOSTAK, *Gossips and telephones*, Discrete Math., 2 (1972), pp. 191–193.
[2] R. T. BUMBY, *A problem with telephones*, this Journal, 2 (1981), pp. 13–18.
[3] D. J. KLEITMAN AND J. B. SHEARER, *Further gossip problems*, Discrete Math., 30 (1980), pp. 151–156.
[4] H. W. LENSTRA, et al. private communication.
[5] D. B. WEST, *A class of solutions to the gossip problem, Part I*, Discrete Math., 39 (1982).
[6] ——, *A class of solutions to the gossip problem, Part III*, Discrete Math., 40 (1982).

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0003 $01.00/0

ON THE COMPUTATION OF THE COMPETITION NUMBER OF A GRAPH*

ROBERT J. OPSUT†

Abstract. This paper examines the problem of recognizing competition graphs (niche overlap graphs), a notion introduced and studied extensively by Cohen [*Food Webs and Niche Space*, Princeton Univ. Press, Princeton, NJ, 1978]. Beginning with an acyclic digraph $F=(V, A)$, define its competition graph $K(F)=(V, E)$ by $(x, y)\in E$ if and only if there exists a w such that $(x, w)\in A$ and $(y, w)\in A$. A graph, G, is a competition graph if there exists an F such that $G=K(F)$. Roberts [Lecture Notes in Mathematics 642, Springer-Verlag, New York, 1978, pp. 477–490] studied recognizing competition graphs and, equivalently, computing an arbitrary graphs competition number, $k(G)$. The competition number, which he showed to be well defined, is the smallest k such that $G\cup I_k$ is a competition graph. In this paper we settle a question posed by Roberts and show that recognizing competition graphs is NP-complete by reducing it to R-CONTENT as defined by Orlin [Nederl. Akad. Wetensch. Proc. Ser. A, 80 (1977), pp. 406–424]. We also give bounds on $k(G)$ in terms of R-Content (G) and compute $k(G)$ for the class of line graphs using a technique similar to that in Roberts.

1. Competition graphs. Suppose F is a food web, a digraph (V, A), where V is a collection of species in an ecosystem and there is an arc from x to y if x preys on y. Following a common assumption in the ecological literature we shall assume F is acyclic. Corresponding to F is a competition graph or a niche overlap graph, $G=(V, E)$, defined as follows: take the vertices of G to be the same species as those of F and connect x and y with an edge if and only if there is a species w such that (x, w) and (y, w) are in A, in other words if and only if x and y have a common prey. In connection with his studies of competition among species, Joel Cohen [1966], [1977], [1978] has studied the following problem. Given a competition graph $G=(V, E)$, what is the smallest k so that we can assign to each vertex x of V a box $B(x)$ (generalized rectangle with sides parallel to the coordinate axes) in Euclidean k-space so that for all $x\neq y$ in V,

$$\{x, y\}\in E \quad \text{iff} \quad B(x)\cap B(y)\neq\varnothing.$$

This smallest k is called the *boxicity* of G. The concept of boxicity was introduced by Roberts [1969], and has since been studied by Gabai [1974], Trotter [1979], Cozzens [1981], and Cozzens and Roberts [1981], [to appear].

Cohen has observed that almost all competition graphs arising from actual ecosystems have boxicity 1. In studying this observation, Roberts [1978] studied the problem of recognizing competition graphs. This is the question we study here. We shall improve upon some of Roberts' results, give a counterexample to one of his conjectures, and show that the problem of recognizing competition graphs is an NP-complete problem. We shall also show that the problem of recognizing competition graphs is tractable for large families of graphs, in particular line graphs.

2. The competition number. Let us note that in an acyclic digraph there is at least one vertex which has no arc out of it (out-degree = 0). Hence a necessary condition for a graph to be a competition graph is that some vertex competes with no other, i.e., there is a vertex which is isolated. It turns out that if enough isolated vertices are

* Received by the editors June 18, 1981, and in revised form August 3, 1981. This research was supported by the U.S. Air Force Office of Scientific Research under contract AFOSR-80-0196A to Rutgers University.

† Energy Information Administration, Department of Energy, Washington, DC 20461.

added to any graph, it becomes a competition graph. Therefore graphs of arbitrarily large boxicity are competition graphs and hence Cohen's finding becomes even more significant.

To see that adding isolated vertices is sufficient for any graph $G=(V,E)$, let $e=|E|$. We build a food web F on $V\cup\{x_\alpha:\alpha\in E\}$ by orienting an arc from the endpoints a and b of α to x_α. Then it is easy to see that $G\cup I_e$ is the competition graph for F, where $G\cup I_e$ means G plus e isolated vertices. Following Roberts [1978] we can now define the competition number of a graph, $k(G)$, as the smallest k such that $G\cup I_k$ is a competition graph. The problem of characterizing the class of competition graphs reduces to finding $k(G)$ for all graphs. Roberts was able to give some results on competition numbers. We restate some of them here. Let $n(G)=|V(G)|$ and $e(G)=|E(G)|$.

PROPOSITION 1. *If G is a graph without triangles (cliques of size 3), then $k(G)\geqq e(G)-n(G)+2$.*

A graph G is called a *rigid circuit graph* if G does not have Z_n, a circuit of length n, $n>3$, as a generated subgraph.

PROPOSITION 2. *Every rigid circuit graph has $k(G)\leqq 1$ with equality if and only if G has no isolated vertices.*

COROLLARY 3. *Every interval graph (a graph of boxicity $\leqq 1$) has $k(G)\leqq 1$.*

PROPOSITION 4. *If G is connected, $n(G)>1$, and G has no triangles, then $k(G)=e(G)-n(G)+2$.*

Roberts also developed a heuristic algorithm which gives a bound, m, on $k(G)$ by constructing a food web whose competition graph is $G\cup I_m$. The algorithm works on an ordering $P=\langle v_1,v_2,\cdots,v_n\rangle$ of the vertices of G. He conjectured that this bound, minimized over all orderings of the vertices, was sharp.

We improve upon Proposition 1 in § 3 and obtain other useful bounds. In § 4, we describe Roberts' algorithm and then present a counterexample to his conjecture. Section 5 deals with the intractability of computing $k(G)$ while in § 6 we show that for line graphs $k(G)$ is easily calculated. (The line graph, G, for a graph G^* has vertex set $E(G^*)$ with two vertices in G adjacent if and only if the corresponding edges in G^* have a vertex in common.)

3. Bounds. In this section, we derive simple bounds on $k(G)$ which will be useful later. We improve upon Roberts' result as stated in Proposition 1. Let $i(G)$ be defined as the least number of cliques which cover the edges of G. Then for triangle-free graphs $e(G)=i(G)$ so we could write the inequality as $k(G)\geqq i(G)-n(G)+2$. In this form we will show that the proposition remains true for arbitrary graphs. The proof is similar to the one Roberts gave for Proposition 1.

PROPOSITION 5. *For any graph G, $k(G)\geqq i(G)-n(G)+2$.*

Proof. Let $n=n(G)$, $i=i(G)$, and $k=k(G)$. Suppose that $G\cup I_k$ is a competition graph with corresponding food web F. According to Corollary 10.1a of Harary, Norman, and Cartwright [1965], we can assign the integers $1,2,\cdots,n+k$ to the $n+k$ vertices of F so that every vertex gets a different integer and every arc goes from a lower number to a higher number. In particular, it follows that the vertex labeled 1 has no incoming arcs and the vertex labeled 2 has at most one incoming arc. Consider the set $P=\{3,4,\cdots,n+k\}$ and for each $j\in P$ the set $K_j=\{x:(x,j)\in A(F)\}$. Then, since $G\cup I_k$ is the competition graph for F, each K_j is a clique of G. Furthermore, $\cup_{j\in P}E(K_j)$ must cover $E(G)$. Hence $i(G)\leqq|P|=n+k-2$ or $k(G)\geqq i(G)-n(G)+2$. Q.E.D.

Roberts went on to prove that if G is connected, triangle-free and nontrivial

($|V(G)|>1$) then the bound is sharp. Unfortunately this result does not hold in the general case as can be seen by taking G to be K_n, the complete graph on n vertices. For $k(G)=1$ while $i(G)-n(G)+2=1-n+2=3-n$.

We can also get an upper bound on $k(G)$ involving $i(G)$ by refining the method we used to show that $k(G)$ is well defined.

PROPOSITION 6. *For any graph* G, $k(G)\leqq i(G)$.

Proof. Let $i=i(G)$ and $K_1, K_2, \cdots, K_i$ be a collection of cliques which cover $E(G)$. Build a food web F on the vertices of $G\cup I_i$ as follows: let the additional isolated vertices be labeled $x_j, j=1, \cdots, i$ and include an arc from the vertices in K_j to x_j. Then F is acyclic and $G\cup I_i$ is the competition graph for F. So by the definition of $k(G)$, $k(G)\leqq i=i(G)$. Q.E.D.

For any graph H, $\theta(H)$ is defined to be the smallest number of cliques that cover the vertices of H. Let the open neighborhood of v in G be denoted by

$$N(v)=\{x\in V(G): \{x, v\}\in E(G)\}.$$

PROPOSITION 7. *For any graph,* $k(G)\geqq \min_v \theta[N(v)]$.

Proof. Let $k=k(G)$ and F be a food web such that F is a competition graph for $G\cup I_k$. Let the vertices of I_k be labeled $x_1, x_2, \cdots, x_k$. Consider $F'=F\sim \{x_1, x_2, \cdots, x_k\}$. Since F' is acyclic there exists a vertex, z, such that the outdegree of z in F' is 0. Hence all the arcs from z in F must go to the vertices in I_k. But this implies that $N(z)\subseteq K_1\cup K_2\cup\cdots\cup K_k$ where $K_j=\{w: (w, x_j)\in A(F)\}$. Hence the K_j's form a clique covering of the vertices of $N(z)$. Therefore, $\min_v \theta[N(v)]\leqq\theta[N(z)]\leqq k$. Q.E.D.

4. Roberts' algorithm and a counterexample. As mentioned above, Roberts [1978] developed an algorithm which gives a bound, m, on $k(G)$ by constructing a food web whose competition graph is $G\cup I_m$. The algorithm operates on the ordering of the vertices of G, $P=\langle v_1, v_2, \cdots, v_n\rangle$. Here, we will describe the algorithm, and then show the resulting bound is not sharp. In order to describe the algorithm we will need some notation. $\bar{N}(v)$, the closed neighborhood of v, is defined as:

$$\bar{N}(v)=\{x\in V(G): \{x, v\}\in E(G)\}\cup\{v\}.$$

We will also use $\bar{N}(v)$ for the subgraph generated (induced) by this set of vertices. Additionally, $G\Delta a$ will denote the subgraph generated by vertices of G other than a, less the edges of $\bar{N}(a)$.

The basic idea is that we build up a food web F in stages, one corresponding to each vertex v_j of G. At each stage, we have a list A_j of vertices of G which could be used as prey in the food web. We use up vertices in A_j first, and then add new vertices not in G as prey to account for the competitions of v_j. At the stage corresponding to v_j, we consider a covering of the edges of $\bar{N}(v_j)$ by cliques. Corresponding to each clique K in this covering, we add arcs from vertices in K to a common prey taken from either A_j or added as a new vertex. This accounts for all competitions in K. In later stages the competitions accounted for in $\bar{N}(v_j)$ are ignored. At the last stage, we count up the number of new vertices added. If there were m, we have a food web whose competition graph is $G\cup I_m$.

The details of each step are now outlined.

Step 0. Set $G_1=G$, $A_1=\varnothing$, $j=1$. Let F_1 have vertex set $V_1=V$ and arc set $B_1=\varnothing$.

Step 1. Calculate $i(\bar{N}_j(v_j))=h_j$, where $\bar{N}_j(v_j)$ is the closed neighborhood of v_j in G_j, and let $K_1, \cdots, K_{h_j}$ be an edge covering of $\bar{N}_j(v_j)$ with $v_j\in K_s$ for $s=1, \cdots, h_j$.

Step 2. If $A_j \neq \varnothing$, add arcs from the vertices in K_s, $s = 1, \cdots, h_j$, to a different vertex a_{js} of A_j until either all the cliques $K_1, \cdots, K_{h_j}$ have been accounted for, or you run out of vertices in A_j.

Step 3. If $A_j = \varnothing$ or A_j was exhausted by Step 2, add arcs from the vertices of the remaining cliques, $K_{|A_j|+1}, \cdots, K_{h_j}$, to new isolated vertices, one for each of the remaining cliques. Set $A_{j+1} = \{v_j\}$ and go to Step 5.

Step 4. If A_j was not exhausted by Step 2, let A_{j+1} equal those vertices of A_j not used in Step 2 plus v_j.

Step 5. Let F_{j+1} have vertex set $V_{j+1} = V_j \cup$ all vertices added in Steps 2 and 3, and arc set $B_{j+1} = B_j \cup$ all arcs added in Steps 2 and 3.

Step 6. If $j = n$, output food web F_{n+1}. If not, set $G_{j+1} = G_j \Delta v$, $j = j+1$, and go to Step 1.

An illustration of the algorithm is given in Fig. 1.

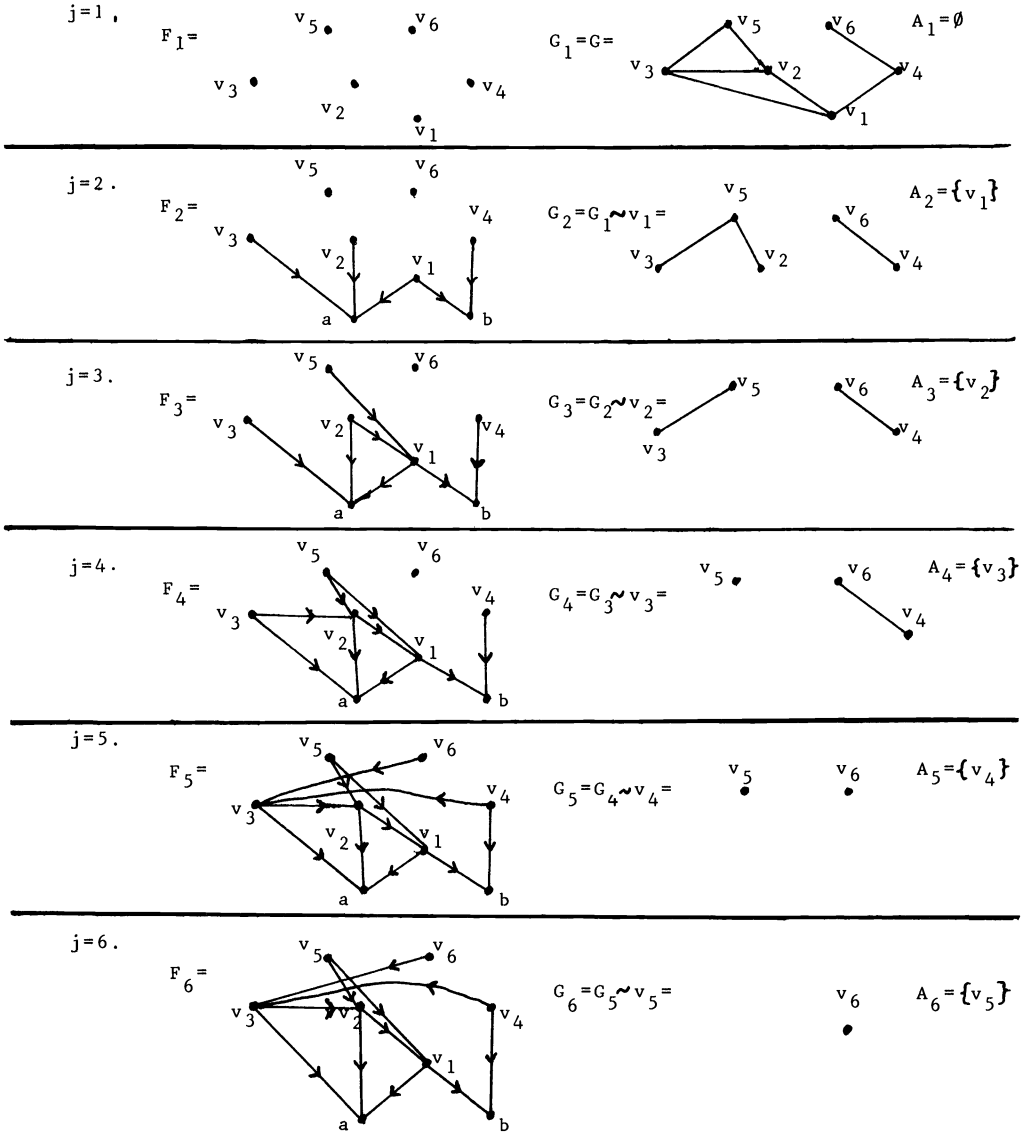


FIG. 1

The proof that the construction does produce a food web whose competition graph is $G \cup I_m$, where m is the number of vertices added, can be found in Roberts [1978]. Let us denote the number, m, of additional isolated vertices produced by the algorithm by $m(G, P)$. Then clearly $k(G) \leqq m(G, P)$. Roberts conjectured that if we looked at all the possible orderings of the vertices of G and let $m(G)$ be the least value of $m(G, P)$, that $m(G)$ would be equal to $k(G)$. We now produce a counter-example. Consider G and F as in Fig. 2. $G \cup I_1$ is the competition graph for F and hence $k(G) = 1$.

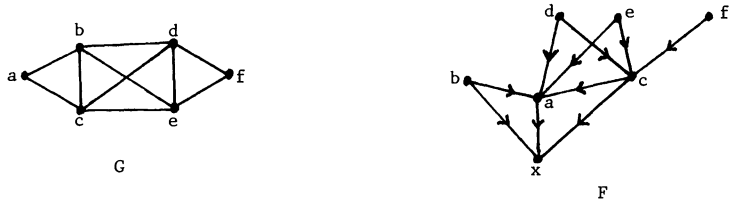


FIG. 2

However, no matter which ordering we choose for the vertices of G, $m(G, P) \geqq 2$ and so $m(G) \geqq 2$. To show $m(G, P) \geqq 2$ for any ordering, suppose $m(G, P) \leqq 1$. Then since $i[\bar{N}(b)] = i[\bar{N}(c)] = i[\bar{N}(d)] = i[\bar{N}(e)] = 2$ we must begin with a or f. Without loss of generality let us begin with $a_1 = a$. We must add one isolated vertex to cover the competitions of a, A_2 becomes $\{a\}$, and $G_2 = G\Delta a$ is shown in Fig. 3. Since $i[\bar{N}_2(d)] = i[\bar{N}_2(e)] = 3$, we must choose b, c, or f to avoid adding any more new vertices. Again without loss of generality choose $a_2 = b$. Then a is used to account for $\bar{N}_2(b)$, $A_3 = \{b\}$ and $G_3 = G_2\Delta b$ is shown in Fig. 4. However, now, no matter which vertex we choose for a_3, $i[\bar{N}_3(a_3)] = 2$ and so the algorithm requires that we add another vertex. Hence $m(G, P) \geqq 2$ for any ordering P.

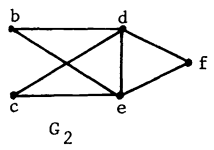


FIG. 3

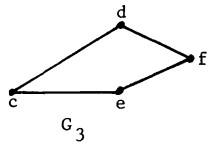


FIG. 4

5. NP-completeness. In this section we show that the problem of determining whether a graph is a competition graph is NP-complete. This result implies that there is little hope for finding an efficient (polynomially bounded) algorithm for this problem or for computing $k(G)$ for G an arbitrary graph. For a definition of NP-completeness and its ramifications, see Garey and Johnson [1979]. The proof hinges on the fact that deciding if for a given graph, G, $i(G) \leqq i$ is NP-complete. This was shown by Orlin [1976], who called $i(G)$ the R-content of a graph, and by Kou, Stockmeyer and Wong [1978]. Let us call the decision problem of recognizing a competition graph, COMPETITION, and the problem of deciding if $i(G) \leqq i$, R-CONTENT.

THEOREM 8. COMPETITION *is* NP-*complete.*

Proof. It is clear that COMPETITION is in NP since given an acyclic digraph, F, on $V(G)$ we can check in polynomial time whether G is the competition graph for F.

To show that COMPETITION is NP-complete, we reduce R-CONTENT to it. Given a graph $G = (V, E)$ for which we are to decide if $i(G) \leqq i$, consider a new graph

$G' = G \cup |V| \cdot K_{2,3} \cup I_{i+2}$. (The unions are disjoint, $K_{2,3}$ is the complete bipartite graph with a bipartition of 2 and 3 vertices, and $n \cdot H$ means n disjoint copies of H.)

Our claim is that $i(G) \leqq i$ if and only if G' is a competition graph. Hence if we had a polynomially bounded algorithm for COMPETITION, we could derive one for R-CONTENT.

To see that if $i(G) \leqq i$, then G' is a competition graph, form an acyclic digraph F on $V(G')$ as follows.

Denote the vertices of G by x_j, $j = 1, \cdots, n$ and label the vertices of the jth copy of $K_{2,3}$ as in Fig. 5. Let $K_1, K_2, \cdots, K_i$ be a clique covering of $E(G)$ which exists since $i(G) \leqq i$. Denote the vertices of I_{i+2} as $w_1, w_2, \cdots, w_i, a_0, c_0$.

Now define the arcs of F as

$$A = \bigcup_{j=1}^{|V|} \{(a_j, b_j), (a_j, a_{j-1}), (a_j, x_j), (b_j, d_j), (b_j, e_j), (b_j, c_{j-1}), (c_j, b_j), (c_j, d_j), (d_j, e_j), (d_j, a_{j-1}), (e_j, c_{j-1}), (e_j, x_j)\}$$

$$\bigcup_{s=1}^{i} \bigcup_{x_j \in K_s} \{(x_j, w_s)\}.$$

One can easily check that F is acyclic and that G' is the competition graph for F.

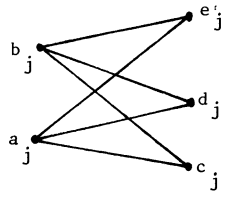


FIG. 5

To see that if G' is a competition graph then $i(G) \leqq i$, note that $i(G') = i(G) + |V| \cdot i(K_{2,3})$ and that $i(K_{2,3}) = 6$. If G' is a competition graph then by the definition of $k(G')$, $k(G') = 0$. By Proposition 5,

$$\begin{aligned} 0 &\geqq i(G') - |V(G')| + 2 \\ &= [i(G) + |V| \cdot i(K_{2,3})] - [|V| + |V| \cdot |V(K_{2,3})| + i + 2] + 2 \\ &= i(G) + 6|V| - |V| - 5|V| - i - 2 + 2 \\ &= i(G) - i \end{aligned}$$

or $i(G) \leqq i$. Q.E.D.

COROLLARY 9. *The computation of $k(G)$ for arbitrary G is* NP*-hard.*

Proof. Determining whether a graph G is a competition graph is equivalent to asking if $k(G) = 0$. Q.E.D.

6. Line graphs. Despite the result of Corollary 9, the computation of $k(G)$ is tractable for large families of graphs. We have seen simple formulas for rigid circuit graphs and connected, triangle-free graphs. In this section we show that when G is a line graph, the computation of $k(G)$ is also tractable. For the proof we will need some classical results whose proofs can be found in Harary [1969]. A claw is a $K_{1,3}$, and "claw-free" means a graph does not have a claw as a generated subgraph.

LEMMA 10. *For each vertex, v, of a line graph, $\theta[N(v)] \leqq 2$. In particular, a line graph is claw-free.*

LEMMA 11. *Any generated (induced) subgraph of a line graph is a line graph.*

THEOREM 12. *If G is a line graph, then $k(G) \leqq 2$, with equality if and only if for all vertices, v, of $V(G)$, $\theta[N(v)] = 2$.*

Proof (by induction on $|V(G)|$). It is clear that when $|V(G)| = 1$, then $k(G) \leqq 1$. So let us assume that $|V(G)| = n$ and that for all line graphs on fewer vertices the theorem holds.

Case 1. There exists a vertex, v, such that $\theta[N(v)] = 0$, i.e., v is an isolated vertex. Then $G \sim v$ (the subgraph generated by all the vertices except v) is also a line graph and so by the induction hypothesis $k(G \sim v) \leqq 2$. Let F^* be a food web on $V(G \sim v) \cup I_2$ where $I_2 = \{a, b\}$ such that $(G \sim v) \cup I_2$ is the competition graph for F^*. Let F^* be the food web on $V(G) \cup \{b\}$ with the following arcs:

$$A(F) = A(F^*) \sim \{(x, a): (x, a) \in A(F^*)\} \cup \{(x, v): (x, a) \in A(F^*)\},$$

i.e., replace a with v. Then it is clear that F is acyclic and $G \cup I_1$ is the competition graph for F. So $k(G) \leqq 1$.

Case 2. There exists a vertex, v, such that $\theta[N(v)] = 1$.

Let $N(v) = \{x_1, x_2, \cdots, x_s\}$. ($s \neq 0$ since $\theta[N(v)] \neq 0$.) Since G is claw-free, $\theta[N(x_j) \sim N(v)] \leqq 1$. For otherwise if $y, z \in N(x_j)$, $y, z \notin N(v)$ and $\{y, z\} \notin E(G)$, then x_j, y, z, v form a claw in G. Now let $G' = G \sim \{v, x_1, x_2, \cdots, x_{s-1}\}$. Then by Lemma 11, G' is a line graph and in G', $\theta[N(x_s)] \leqq 1$. Hence by the induction hypothesis there exists a food web F^* such that $G' \cup I_1$ is the competition graph for F^*. Let $I_1 = \{a\}$. Construct a food web F on $V(G \cup I_1)$ by defining

$$\begin{aligned} A(F) = A(F^*) \sim \{(w, a): (w, a) \in A(F^*)\} \cup \{(w, x_{s-1}): (w, a) \in A(F^*)\} \\ \cup \bigcup_{j=2}^{s-1} \{(x_j, x_{j-1}), (y, x_{j-1}): y \in N(x_j) \sim \bar{N}(v)\} \\ \cup \{(x_1, v)\} \cup \{(y, v): y \in N(x_1) \sim \bar{N}(v)\} \\ \cup \{(v, b)\} \cup \{(x_j, b): x_j \in N(v)\} \end{aligned}$$

i.e., replace a with x_{s-1}, use x_{j-1} to account for the competitions of x_j not in $\bar{N}(v)$, use v to account for the competitions of x_1 not in $\bar{N}(v)$ and add vertex b to account for the competitions in $\bar{N}(v)$. Hence $G \cup I_1$ is the competition graph of F and since F is acyclic, $k(G) \leqq 1$.

Case 3. For each v, $\theta[N(v)] = 2$.

Let G^* be such that G is the line graph of G^*.

Choose any vertex v and let v correspond to the edge $\{\alpha, \beta\}$ in G^*. Let $N(v) = \{x_1, x_2, \cdots, x_s, y_1, \cdots, y_p\}$ where x_j corresponds to edge $\{\alpha, \delta_j\}$ and y_r corresponds to $\{\beta, \tau_r\}$. Then $N(x_j) \sim \bar{N}(v)$ is the clique $K_j = \{w: w = \{\delta_j, \sigma_t\}$ for some $\sigma_t\}$. In particular in $G' = G \sim \{v, x_1, \cdots, x_{s-1}\}$, $\theta[N(x_s)] \leqq 1$. Since G' is a line graph, the induction hypothesis implies there is a food web F^* whose competition graph is $G' \cup I_1$. Let the isolated vertex be labeled a. Then we can build a food web F on $V(G \cup I_2)$ where $I_2 = \{b, c\}$ with the following arcs:

$$\begin{aligned} A(F) = A(F^*) \sim \{(w, a): (w, a) \in A(F^*)\} \cup \{(w, x_{s-1}): (w, a) \in A(F^*)\} \\ \bigcup_{j=2}^{s-1} \{(w, x_{j-1}): w = x_j \text{ or } w \in K_j\} \end{aligned}$$

$$\cup\{(w, v): w = x_1 \text{ or } w \in K_1\}$$
$$\cup\{(x_j, b): j = 1, \cdots, s\}$$
$$\cup\{(y_r, c): r = 1, \cdots, p\} \cup \{(v, b), (v, c)\}.$$

In other words, replace a with x_{s-1}, have $\{x_j\} \cup K_j$ feed on x_{j-1}, $\{x_1\} \cup K_1$ feed on v, and account for the competitions in $\bar{N}(v)$ with b and c.

It is easy to see that F is acyclic and $G \cup I_2$ is the competition graph for F, so $k(G) \leqq 2$.

By Proposition 7, since each vertex of G has $\theta[N(v)] = 2$, $k(G) \geqq 2$ and so $k(G) = 2$. Q.E.D.

Theorem 12 gives us a simple procedure to calculate $k(G)$ for line graphs. Those without a simplicial vertex, i.e., $\theta[N(v)] = 2$ for all v, have $k(G) = 2$. The other two cases can be distinguished as follows. $k(G) = 1$ if and only if either (a) there is no isolated vertex and there is a simplicial vertex in G, or (b) there is a unique isolated vertex, w, and there is no simplicial vertex in $G \sim w$. $k(G) = 0$ if and only if either (a) there is a unique isolated vertex, w, and there is a simplicial vertex in $G \sim w$, or (b) there are two isolated vertices.

7. Further questions.

1. While we have given a counterexample to Roberts' conjecture that $k(G) = m(G)$, is there another procedure that reduces computing $k(G)$ to computing $i(G)$?

2. Cases 1 and 2 in the proof of Theorem 11 require only that $\theta[N(v)] \leqq 2$ for all v in G. Can the result be strengthened to state that if G is any graph with $\theta[N(v)] \leqq 2$ for all vertices v of G, then $k(G) \leqq 2$ with equality if and only if $\theta[N(v)] = 2$ for all v? The author conjectures that this is true.

3. Is there a characterization of food webs whose competition graphs have boxicity $\leqq 1$? This might shed some light on Cohen's findings.

REFERENCES

J. E. COHEN [1966], *A Model of Simple Competition*, Harvard Univ. Press, Cambridge, MA.

———, [1977], *Food webs and the dimensionality of trophic niche space*, Proc. Nat. Acad. Sci. U.S.A., 74, pp. 4533–4536.

———, [1978], *Food Webs and Niche Space*, Princeton Univ. Press, Princeton, NJ.

M. B. COZZENS [1981], *The NP-completeness of the boxicity of a graph*, in preparation.

———, [to appear], *Computing the boxicity of a graph by covering its complement by cointerval graphs*, Discrete Applied Math.

M. B. COZZENS AND F. S. ROBERTS [1981], *Meteoric $(2k+1)$-tuples and the boxicity of a graph*, mimeographed, Rutgers Univ., New Brunswick, NJ.

H. GABAI [1974], *Bounds for the boxicity of a graph*, mimeographed, York College, City Univ. of New York.

M. R. GAREY AND D. S. JOHNSON [1979], *Computers and Intractability, A Guide to the Theory of NP Completeness*, W. H. Freeman, San Francisco.

F. H. HARARY [1969], *Graph Theory*, Addison-Wesley, Reading, MA.

F. HARARY, R. Z. NORMAN AND D. CARTWRIGHT [1965], *Structural Models: An Introduction to the Theory of Directed Graphs*, John Wiley, New York.

L. T. KOU, L. J. STOCKMEYER AND C. K. WONG [1978], *Covering edges by cliques with regard to keyword conflicts and intersection graphs*, Comm. ACM, 21, pp. 135–138.

J. ORLIN [1977], *Contentment in graph theory: Covering graphs with cliques*, Nederl. Akad. Wetensch. Proc. Ser. A, 80, pp. 406–424.

F. S. ROBERTS [1969], *On the boxicity and cubicity of a graph*, in Recent Progress in Combinatorics, Academic Press, New York, pp. 301–310.

———, [1976], *Discrete Mathematical Models, With Applications to Social, Biological and Environmental Problems*, Prentice-Hall, Englewood Cliffs, NJ.

———, [1978], *Food webs, competition graphs and the boxicity of ecological phase space*, in Theory and Applications of Graphs—In America's Bicentennial Year, Lecture Notes in Mathematics 642, Y. Alavi and D. Lick, eds., Springer-Verlag, New York, pp. 477–490.

W. T. TROTTER, JR. [1979], *A forbidden subgraph characterization of Roberts' inequality for boxicity*, Discrete Math., 28, pp. 303–314.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0004 $01.00/0

ALGORITHMS FOR TESTING THE DIAGONAL SIMILARITY OF MATRICES AND RELATED PROBLEMS*

GERNOT M. ENGEL† AND HANS SCHNEIDER‡

Abstract. A simple algorithm is presented for testing the diagonal similarity of two square matrices with entries in a field. Extended forms of the algorithm decide various related problems such as the simultaneous diagonal similarity of two families of matrices, the existence of a matrix in a subfield diagonally similar to a given matrix, the existence of a unitary matrix similar to a given complex matrix, and the corresponding problems for diagonal equivalence in place of diagonal similarity. The computational complexity of our principal algorithm is studied, programs and examples are given. The algorithms are based on the existence of a canonical form for diagonal similarity. In the first part of the paper theorems are proved which establish the existence of this form and which investigate its properties.

1. Introduction. In this paper we present a simple algorithm for testing the diagonal similarity of two square matrices with entries in a field $\mathbb{F}$. Extended forms of our algorithm decide the simultaneous diagonal similarity of two families of matrices, the existence of a matrix in a subfield diagonally similar to a given matrix and, if $\mathbb{F}$ is the real or complex field, the existence of a real orthogonal or unitary matrix diagonally similar to a given matrix. Another modification of our algorithm tests the diagonal equivalence of two rectangular matrices. There exist extensions for diagonal equivalence which correspond to the extensions described above in the case of diagonal similarity.

After the appropriate definitions (§ 2), we develop the theory on which our algorithm is based (§ 3 and § 4). We show that for $A \in \mathbb{F}^{nn}$, the set of $n \times n$ matrices with elements in $\mathbb{F}$, there exists a canonical form for diagonal similarity. We denote this form by A_F, since it depends on a choice of a spanning forest F for the graph $G(A)$ of A considered as an undirected multigraph. Further, we give a simple construction for a diagonal matrix X such that $XAX^{-1} = A_F$ and we write $X = X(A, F, U)$ since X also depends on a choice of a set of representatives U for the connected components of F or $G(A)$. Thus, for $A, B \in \mathbb{F}^{nn}$, the matrices A and B are diagonally similar if and only if $G(A) = G(B)$ and $A_F = B_F$ or, equivalently, H_F is a $\{0, 1\}$ matrix where $H = A \oslash B$ is the Hadamard quotient defined in [1] or § 3. Thus we have the following simple procedure to test diagonal similarity of A and B:

(1) Check whether $G(A) = G(B)$.
(2) If so, choose a spanning forest F for $G(A)$ and a set U of representatives for the connected components of $G(A)$.
(3) For $H = A \oslash B$, compute $X = X(H, F, U)$.
(4) Check whether $XHX^{-1} \in \{0, 1\}^{nn}$.

A detailed description of the algorithm and a study of its computational complexity is given in § 5. In § 6 we briefly indicate applications which are more fully described in our technical report with the same title as this paper.

The relationship between cyclic products and diagonal similarity which is crucial to our theory can be traced back as far as Fiedler–Pták [4]. Theorems with proofs on which algorithms may be based are given in [6], e.g. Theorem 3.17, though no actual

* Received by the editors January 30, 1981, and in revised form July 28, 1981.

† 2154 Pinar Place, Del Mar, California 92014.

‡ Mathematics Department, University of Wisconsin, Madison, Wisconsin 53706. The research of this author was supported in part by the National Science Foundation under grants MCS 78-01087 and MCS 80-26132 and by the Deutsche Forschungsgemeinschaft (DFG) while he was a Visiting Professor at the Mathematisches Institut der Universität Würzburg, D-8700 Würzburg, West Germany.

algorithm is to be found in that paper. The cycles used in these papers are of a restricted type which occur in the evaluation of determinants; i.e., an arc (i, j) is traversed only from i to j. In view of this, unless there is an irreducibility condition on the matrix, any algorithm based on these results requires the determination of the Frobenius block form of the matrix. For the special problem of diagonal similarity to a unitary matrix an interesting algorithm of this type is to be found in Berman–Parlett–Plemmons [1]. The use of general cycles to prove results on diagonal similarity occurs in [6]. Though the proofs in that paper are geometric and existential, it is these features which allow us here to develop constructive proofs and algorithms which do not require the Frobenius block form. The corresponding tool is a spanning forest of an undirected multigraph, which has already been mentioned and which is simple to compute. Thus our algorithm appears to have computational advantages.

2. Definitions.

DEFINITION 2.1. Formally, a (simple, directed) graph G is a pair $G = (I, E)$ of finite sets with $E \subseteq I \times I$. The elements of I are called the *vertices* of G, and the elements of E the *arcs* of G. We represent graphs in the usual way, see, e.g., Fig. 1, where $e_1 = (1, 2)$, etc.

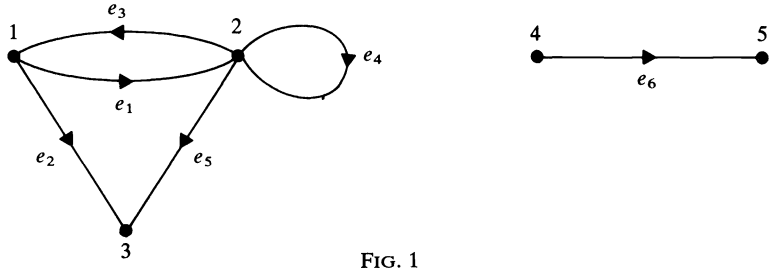


FIG. 1

Since this graph will be used as an example several times, we shall call it G^*. Although in Fig. 1 we use arrows to represent arcs, we give the symbols $i \to j$ and $j \leftarrow i$ somewhat different meanings in the text. A *link* in G is a triple $\lambda = (i, j, \varepsilon)$ where $(i, j) \in E$ and $\varepsilon = \pm 1$.

If $\varepsilon = +1$, $(\varepsilon = -1)$ we call i the *start* (*end*) and j the *end* (*start*) of λ. Intuitively, we may consider $(i, j, +1)$ as the arc (i, j) traversed from i to j, and $(i, j, -1)$ as the same arc, traversed from j to i. Thus it is natural to represent $(i, j, +1)$ by $i \to j$ and $(i, j, -1)$ by $j \leftarrow i$.

A chain in G is a sequence $\alpha = (\lambda_1, \cdots, \lambda_s)$ of links in G for which the end of λ_p is the start of λ_{p+1}, $p = 1, \cdots, s-1$. The start i of α is the start of λ_1, the *end* j of α is the end of λ_s. We also say that α is a chain from i to j. Our notation for links is immediately extended to chains, as we illustrate by means of examples from the graph G^* of Fig. 1:

Thus

$$\alpha = 3 \leftarrow 1 \to 2 \to 2, \qquad \beta = 3 \leftarrow 1 \leftarrow 2 \to 2$$

respectively stand for the chains

$$\alpha = ((1, 3, -1), (1, 2, +1), (2, 2, +1)),$$

$$\beta = ((1, 3, -1), (2, 1, -1), (2, 2, +1))$$

from 3 to 2. Observe that α traverses the arc $(1, 2)$, while β traverses the arc $(2, 1)$. Thus the concept of chain formalizes the notion of putting a pencil on a vertex of a graph represented as in Fig. 1 and moving it in or against the direction of a sequence arcs to another vertex.

Let $\alpha = (\lambda_1, \cdots, \lambda_s)$ be a chain in G. We call α a *simple chain* if the starts of $\lambda_1, \cdots, \lambda_s$ are pairwise distinct. We call α a *closed chain* if the start and end of α coincide. A simple closed chain is called a *cycle.*

If $\alpha = (\lambda_1, \cdots, \lambda_s)$ and $\beta = (\lambda_{s+1}, \cdots, \lambda_{s+t})$ are chains such that the end of α coincides with the start of β then $\alpha\beta$ denotes the chain $(\lambda_1, \cdots, \lambda_{s+t})$. If $\lambda = (i, j, +1)$ is a link then $\lambda^{-1} = (i, j, -1)$ and if α is the chain above then $\alpha^{-1} = (\lambda_s^{-1}, \cdots, \lambda_1^{-1})$. It will also be convenient to regard $\varnothing$ as the empty chain from any vertex to itself.

DEFINITION 2.2. A *subgraph* of $G = (I, E)$ is a graph $G' = (I', E')$ such that $I' \subseteq I$, $E' \subseteq E$. We write $G' \subseteq G$. Let $F = (I', E')$. We call F a *forest* if F has no cycles. A maximal forest F contained in G is called a *spanning forest,* viz. F is a forest and if F' is a forest for which $F \subseteq F' \subseteq G$ then $F' = F$. It is well known that every graph $G = (I, E)$ has a spanning forest $F = (I', E')$ and that $I' = I$.

DEFINITION 2.3. A graph $G = (I, E)$ is *connected* if for each pair of vertices $\{i, j\}$ there is a chain in G from i to j. (Observe that a graph with a single vertex is connected since $\varnothing$ is a chain). A maximal connected subgraph of G is called a *component* of G. A connected forest is called a *tree,* a connected spanning forest of G is called a *spanning tree* of G. The components of a forest are trees.

For example, a component of the graph of G^* of Fig. 1 is $G_2^* = (\{4, 5\}, \{e_6\})$. A spanning forest of this graph has components $G_1^* = (\{1, 2, 3\}, \{e_1, e_2\})$ and G_2^*.

Let G be a graph with components $G_1, \cdots, G_t$. If i_p is a vertex of G_p, $p = 1, \cdots, t$ we call $U = \{i_1, \cdots, i_p\}$ a *set of representatives* for G. If F is a spanning forest for G, then U is also a set of representatives for F. For example $U^* = \{1, 4\}$ is a set of representatives for (G^* and) the spanning forest F^*.

If F is a tree and i, j are vertices in F, then it is easy to see that there is a unique simple chain in F from i to j. If G is a graph, F a spanning tree for G and $e = (i, j)$ an arc of G which is not in F, (write $e \in G \backslash F$) then there is a unique cycle $\gamma = (\lambda, \lambda_1, \cdots, \lambda_s)$ such that $\lambda = (i, j, +1)$ and $(\lambda_1, \cdots, \lambda_s)$ is a chain in F. We call this cycle the *canonical cycle for e with respect to F.*

3. Main theoretical results. Subsequently, $\mathbb{F}$ will be a field and $\mathbb{F}^{nn}$ the set of all $(n \times n)$ matrices with entries in $\mathbb{F}$.

DEFINITION 3.1. Let $A, B \in \mathbb{F}^{nn}$. Then A is *diagonally similar* to B if there exists a (nonsingular) diagonal matrix X in $\mathbb{F}^{nn}$ for which $XAX^{-1} = B$.

DEFINITION 3.2. Let $A \in \mathbb{F}^{nn}$. Let $\langle n \rangle = \{1, \cdots, n\}$. We define the graph $G(A) = (I, E)$ of A thus:

$$I = \langle n \rangle, \qquad (i, j) \in E \quad \text{if } a_{ij} \neq 0, \quad i, j = 1, \cdots, n.$$

DEFINITION 3.3. Let $A \in \mathbb{F}^{nn}$ and let $\alpha = (\lambda_1, \cdots, \lambda_s)$ be a chain in $G(A)$, where $\lambda_p = (i_p, j_p, \varepsilon_p)$, $p = 1, \cdots, s$. Then the *chain product* $\pi_\alpha(A)$ is defined by

$$\pi_\alpha(A) = a_{i_1 j_1}^{\varepsilon_1} a_{i_2 j_2}^{\varepsilon_2} \cdots a_{i_s j_s}^{\varepsilon_s}.$$

If $\varnothing$ is the empty chain, $\pi_\varnothing(A) = 1$. If $\alpha\beta$ is defined then $\pi_{\alpha\beta}(A) = \pi_\alpha(A)\pi_\beta(A)$ and $\pi_{\alpha^{-1}}(A) = \pi_\alpha(A)^{-1}$. If α is a cycle we call $\pi_\alpha(A)$ a cycle product, etc.

Example 3.4. Let

$$A^* = \begin{bmatrix} 0 & 1 & 2 & 0 & 0 \\ 3 & 4 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then $G(A) = G^*$.

Consider the chains $\alpha = 3 \leftarrow 1 \rightarrow 2 \rightarrow 2$, and $\beta = 3 \leftarrow 1 \leftarrow 2 \rightarrow 2$. Then $\pi_\alpha(A) = a_{13}^{-1} a_{12} a_{22}$ and $\pi_\beta(A) = a_{13}^{-1} a_{21}^{-1} a_{22}$.

DEFINITION 3.5. Let $A \in \mathbb{F}^{nn}$. Let F be a spanning forest for $G(A) = G$. We define the *canonical form* $A_F = C = C(A, F)$ of A (with respect to F) thus: For $1 \leqq i, j \leqq n$,

$$c_{ij} = \begin{cases} 0 & \text{if } (i,j) \notin G, \\ 1 & \text{if } (i,j) \in F, \\ \pi_\gamma(A) & \text{if } (i,j) \in G \backslash F, \end{cases}$$

where γ is the canonical cycle for (i, j) with respect to F.

DEFINITION 3.6. Let F be a spanning forest for the graph $G(A)$, where $A \in \mathbb{F}^{nn}$. Let $U = \{i_1, \cdots, i_t\}$ be a set of representatives for $G(A)$; cf. Definition 2.3. We define a *transforming matrix* $X = X(A, F, U)$ by $x_j = x_\beta(A)$, where, for j in the component G_p of $G(A)$, we denote by β the unique simple chain in F from $i_p \in U$ to j. (Thus $x_j = 1$ if $j \in U$, for then $\beta = \varnothing$.)

Example 3.7. For the matrix A^* of Example 3.4, and F^* and U^* as in Definition 2.3,

$$A_F^* = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 3 & 4 & \frac{5}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

$$X^* = X(A^*, F, U) \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 6 \end{bmatrix}.$$

Note that $X^* A^* (X^*)^{-1} = A_F^*$. We now prove that this is true in general.

THEOREM 3.8. *Let $A \in \mathbb{F}^{nn}$. Let F be a spanning forest for the graph of $G(A)$ and let U be a set of representatives for $G(A)$. If A_F is the canonical form of A with respect to F, and $X = X(A, F, U)$ is a transforming matrix, then $XAX^{-1} = A_F$.*

Proof. Let $C = A_F$.

(i) If $(i, j) \notin G(A)$, then evidently $c_{ij} = 0$.

(ii) Let $(i, j) = e \in F$, say $e \in F_p$, $1 \leqq p \leqq t$. Let β_i, β_j be the unique simple chains from i_p to i and j respectively. Then either $\beta_j = \beta_i(i \rightarrow j)$ or $\beta_i = \beta_j(j \rightarrow i)$. Hence either $x_j = x_i a_{ij}$ or $x_i = x_j a_{ij}^{-1}$. It follows that $x_i a_{ij} x_j^{-1} = 1$.

(iii) Let $(i, j) \in G \backslash F$. Then the vertices i, j belong to a common component F_p of F. If β_i, β_j are defined as above, then we may write $\beta_i = \delta \beta_i'$ and $\beta_i = \delta \beta_j'$, where the chains β_i' and β_j' have no common link. Hence

$$\pi_\alpha(A) = a_{ij} \pi_{\beta_j'}(A)^{-1} \pi_{\beta_j'}(A) = a_{ij} \pi_{\beta_j}(A)^{-1} \pi_{\beta_i}(A) = x_i a_{ij} x_j^{-1}.$$

The matrix A_F is indeed a canonical form for A under diagonal similarity. This will be shown in the next corollary.

COROLLARY 3.9. *Suppose that $A, B \in \mathbb{F}^{nn}$. Let F be a spanning forest for $G(A)$. Then the following are equivalent.*

(i) *A is diagonally similar to B,*

(ii) *$G(A) = G(B)$ and $A_F = B_F$.*

Proof. (ii) $\Rightarrow$ (i). By Theorem 3.8, A is diagonally similar to A_F and B is diagonally similar to B_F. Hence A is diagonally similar to B.

(i) $\Rightarrow$ (ii). Let A be diagonally similar to B. Evidently $G(A) = G(B)$ and $\pi_\gamma(A) = \pi_\gamma(B)$ for all cycles γ in G. Hence by definition of A_F, it follows that $A_F = B_F$.

We state our next corollary in terms of the *Hadamard quotient* $A \textcircled{\div} B$ of two matrices A, B; cf. [1].

DEFINITION 3.10. Let $A, B \in \mathbb{F}^{nn}$ and suppose that $G(A) = G(B)$. Then the *Hadamard quotient* $H = A \textcircled{\div} B$ is defined by

$$h_{ij} = \begin{cases} a_{ij}/b_{ij} & \text{if } (i,j) \in G(A), \\ 0 & \text{otherwise.} \end{cases}$$

It is clear that $YAY^{-1} = B$ is equivalent to $Y(A \textcircled{\div} B)Y^{-1} \in \{0,1\}^{nn}$. Also, a $\{0,1\}$-matrix is in canonical form. Thus, we obtain our chief theoretical tool as an immediate application of Corollary 3.9.

COROLLARY 3.11. *Let $A, B \in \mathbb{F}^{nn}$. Let F be a spanning forest for $G(A)$, with U a set of representatives for $G(A)$. The following are equivalent.*

(i) *A is diagonally similar to B.*

(ii) *$G(A) = G(B)$ and $A \textcircled{\div} B$ is diagonally similar to a $\{0,1\}$-matrix.*

(iii) *$G(A) = G(B)$ and $(A \textcircled{\div} B)_F$ is a $\{0,1\}$-matrix.*

(iv) *$G(A) = G(B)$ and if $X = X(A \textcircled{\div} B, F, U)$ then $XAX^{-1} = B$.*

Our algorithm is based on the equivalence of (i) and (iv) of the above theorem. It rests on the computation of $X = X(A \textcircled{\div} E, F, U)$ and $X(A \textcircled{\div} B)X^{-1}$. Even though there may be other diagonal matrices Y such that $YAY^{-1} = B$, we emphasize that either $XAX^{-1} = B$ or else A is not diagonally similar to B. We now determine those Y for which $YAY^{-1} = B$.

THEOREM 3.12. *Let $A, B \in \mathbb{F}^{nn}$. Let F be a spanning forest for $G(A)$ with components $F_1, \cdots, F_t$ and let U be a set of representatives for $G(A)$. Let $Y \in \mathbb{F}^{nn}$ be diagonal. The following are equivalent.*

(i) $YAY^{-1} = B$.

(ii) *$G(A) = G(B)$ and, for $i \in F_p$ and $i_p \in U$, $y_i = y_{i_p} x_i$, where $X = X(A \textcircled{\div} B, F, U)$.*

Proof. (ii) $\Rightarrow$ (i). Suppose $j \notin F_p$. Then $a_{ij} = 0$ hence also $b_{ij} = 0$ and so $y_i a_{ij} y_j^{-1} = b_{ij}$. If $j \in F_p$, then $y_i a_{ij} y_j^{-1} = y_{i_p} x_i a_{ij} x_j y_{i_p}^{-1} = x_i a_{ij} x_j^{-1} = b_{ij}$.

(i) $\Rightarrow$ (ii). Evidently $G(A) = G(B)$. Since $i \in F_p$, there exists a simple chain γ from i_p to i. Let $H = A \textcircled{\div} B$. Then $YHY^{-1} = XHX^{-1}$. Hence $y_{i_p} \pi_\gamma(H) y_i^{-1} = \pi_\gamma(YHY^{-1}) = \pi_\gamma(XHX^{-1}) = x_{i_p} \pi_\gamma(H) x_i^{-1}$. But $x_{i_p} = 1$ and $x_i = \pi_\gamma(H)$. Hence $y_i = y_{i_p} x_i$.

COROLLARY 3.13. *Let $A, B \in \mathbb{F}^{nn}$ and suppose that A is diagonally similar to B. Let F be a spanning forest for $G(A)$ and U a set of representatives for $G(A)$. Then $X = X(H, F, U)$, where $H = A \textcircled{\div} B$, is the unique matrix which satisfies $XAX^{-1} = B$ and $x_i = 1$ for $i \in U$.*

The impact of Corollary 3.13 is this. If A is diagonally similar to B, then the matrix X which is given by our algorithm and which satisfies $XAX^{-1} = B$ is in fact independent of the choice of the spanning forest F. Another immediate corollary to Theorem 3.12 is the following result, proved by a different method in [6, Prop. 2.3].

COROLLARY 3.14. *Let $A, B \in \mathbb{F}^{nn}$ and suppose that A is diagonally similar to B. Then the following are equivalent.*

(i) *$YAY^{-1} = B$ implies that $Y = cX(A \textcircled{\div} B, F, U)$ where $0 \neq c \in \mathbb{F}$.*

(ii) *$G(A) = G(B)$ is connected.*

4. Applications.

4.1. Simultaneous diagonal similarity.

DEFINITION 4.1. Let P be an index set, and let $A^{(p)}, B^{(p)} \in \mathbb{F}^{nn}$, for $p \in P$. Then the families $\{A^{(p)} : p \in P\}$, $\{B^{(p)} : p \in P\}$ are *simultaneously diagonally similar* if there is a diagonal matrix $X \in \mathbb{F}^{nn}$, $XA^{(p)}X^{-1} = B^{(p)}$, for all $p \in P$.

Our algorithm can easily be adapted to test for the simultaneous diagonal similarity of finite families of matrices. There is no difficulty in proving the underlying theorem when the index set P is infinite.

DEFINITION 4.2. Let $\{H^{(p)} : p \in P\}$ be a family of matrices in $\mathbb{F}^{nn}$. We call the family *semiconstant* if it satisfies the following condition:

$$\text{For } q, p \in P, \text{ and } 1 \leqq i, j \leqq n \text{ either } h_{ij}^{(p)} = h_{ij}^{(q)} \text{ or } h_{ij}^{(p)} h_{ij}^{(q)} = 0.$$

In this case the *supremum matrix* $S = S(H^{(p)} : p \in P)$ is defined thus:

For $i, j = 1, \cdots, n$,

$$s_{ij} = \begin{cases} h_{ij}^{(p)} & \text{if there exists } p \in P \text{ for which } h_{ij}^{(p)} \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

THEOREM 4.3. *Let $\{A^{(p)} : p \in P\}$, $\{B^{(p)} : p \in P\}$ be two families of matrices in $\mathbb{F}^{nn}$. Then the following are equivalent.*

(i) *The families are simultaneously diagonally similar.*

(ii) (a) $G(A^{(p)}) = G(B^{(p)})$, *for* $p \in P$.

(b) *If* $H^{(p)} = A^{(p)} \oplus B^{(p)}$, *then* $\{H^{(p)} : p \in P\}$ *is a semiconstant family of matrices.*

(c) *Let* $S = S(H^p : p \in P\}$ *be the corresponding supremum matrix, and let* F *be a spanning forest of* $G(S)$. *Then the canonical form* $S_F \in \{0, 1\}^{nn}$.

Proof. (ii) $\Rightarrow$ (i). Let $XSX^{-1} \in \{0, 1\}^{nn}$. Then either $s_{ij} = x_j x_i^{-1}$ or $s_{ij} = 0$, $i, j = 1, \cdots, n$. Hence, for each $p \in P$, $h_{ij}^{(p)} = x_j x_i^{-1}$ or $h_{ij}^{(p)} = 0$, $i, j = 1, \cdots, n$. It follows that $b_{ij}^{(p)} = x_i a_{ij}^{(p)} x_j^{-1}$, $i, j = 1, \cdots, n$.

(i) $\Rightarrow$ (ii). Evidently $G(A^{(p)}) = G(B^{(p)})$, for $p \in P$. By assumption there exists a diagonal $Y \in \mathbb{F}^{nn}$ for which $YA^{(p)}Y^{-1} = B^{(p)}$, for $p \in P$. Hence $YH^{(p)}Y^{-1} = \{0, 1\}^{nn}$, for $p \in P$. Thus either $h_{ij}^{(p)} = y_i y_j^{-1}$ or $h_{ij}^{(p)} = 0$, $1 \leqq i, j \leqq n$. Hence $\{H^{(p)} : p \in P\}$ is semi-constant. Let S be the corresponding supremum matrix. It follows that $YSY^{-1} \in \{0, 1\}$, but then $S_F(YSY^{-1})_F \in \{0, 1\}^{nn}$.

In order to test whether $S_F \in \{0, 1\}^{nn}$, we need merely to construct a transforming matrix $X = X(S, F, U)$. Hence we have an effective test for simultaneous diagonal similarity.

4.2. Diagonal similarity to a matrix with elements in a subgroup. It is easily seen that all our previous results hold when $\mathbb{F}$ is a (multiplicative) Abelian group with 0, viz. $\mathbb{F}\backslash\{0\}$ is an Abelian group and $0c = 0 = c0$ for all $c \in \mathbb{F}$. In our next theorem we shall explicitly assume that $\mathbb{F}$ is an Abelian group with 0 and $\mathbb{F}_1$ will be a subgroup with 0. As an example, $\mathbb{F}$ can be chosen to be a field and $\mathbb{F}_1$ a subfield, e.g., $\mathbb{F}$ is the real field and $\mathbb{F}_1$ the rational field. In another important example $\mathbb{F}$ consists of the reals (rationals) and $\mathbb{F}_1$ of the nonnegative reals (rationals).

THEOREM 4.4. *Let $\mathbb{F}$ be an Abelian group with 0 and let $\mathbb{F}_1$ be a subgroup with 0. Let $A \in \mathbb{F}^n$. Then the following are equivalent.*

(i) *For some diagonal matrix $X \in \mathbb{F}^{nn}$, $XAX^{-1} \in \mathbb{F}_1^{nn}$.*

(ii) $A_F \in \mathbb{F}_1^{nn}$.

Proof. (ii) $\Rightarrow$ (i). Trivial, since $A_F = (XAX^{-1})_F$.

(i) $\Rightarrow$ (ii). Suppose that $XAX^{-1} \in \mathbb{F}_1^{nn}$. Then for every cycle γ of $G(A)$ we have $\pi_\gamma(A_F) = \pi_\gamma(A) = \pi_\gamma(XAX^{-1}) \in \mathbb{F}_1$. Since $0, 1 \in \mathbb{F}_1$, it follows that $A_F \in \mathbb{F}_1^{nn}$.

At this point it is appropriate to state an easy result that will be used in § 4.3. With the notation of Theorem 4.4, we observe that $H \in \mathbb{F}_1^{nn}$ implies that $X = X(H, F, U) \in \mathbb{F}_1^{nn}$. The rest of the proof follows from Corollary 3.11.

THEOREM 4.5. *Let $\mathbb{F}$ and $\mathbb{F}_1$ be defined as in Theorem 4.4. Let $A, B \in \mathbb{F}^{nn}$. Then the following are equivalent.*

(i) *There is a diagonal matrix* $Y \in \mathbb{F}_1^{nn}$ *for which* $YAY^{-1} = B$.
(ii) (a) $G(A) = G(B)$.
(b) $(A \circledcirc B)_F \in \{0, 1\}^{nn}$.
(c) $A \circledcirc B \in \mathbb{F}_1^{nn}$.
(iii) *Conditions* (ii) (a), (b) *hold and*
(c) $X(A \circledcirc B, F, U) \in \mathbb{F}_1^{nn}$ *where, as usual,* F *is a spanning forest for* $G(A)$ *and* U *a set of representatives for* $G(A)$.

Thus our algorithm tests, for example, if two real matrices are similar by means of a diagonal matrix with positive diagonal entries.

4.3. Diagonal similarity to a unitary matrix. We now prove results for real or complex matrices related to those in [1]. We shall give necessary and sufficient conditions for a complex matrix to be diagonally similar via a complex diagonal similarity to a unitary matrix and for a real matrix to be diagonally similar via a real similarity to an orthogonal matrix. Our results can be stated as one theorem, since a unitary matrix with real entries is of course orthogonal. We call a matrix Y *nonnegative* if all its entries are nonnegative and we write $Y \geqq 0$.

THEOREM 4.6. *Let* $\mathbb{F} = \mathbb{R}$ *or* $\mathbb{F} = \mathbb{C}$, *and let* $A \in \mathbb{F}^{nn}$. *Then the following are equivalent.*

(i) *There exists a unitary matrix* $B \in \mathbb{F}^{nn}$ *such that* A *and* B *are diagonally similar.*
(ii) (a) A *is nonsingular.*
(b) $YA^{-1}Y^{-1} = A^*$, *for some diagonal* $Y \in \mathbb{F}^{nn}$ *where* $Y \geqq 0$.

Proof. (i) $\Rightarrow$ (ii). Let $B = ZAZ^{-1}$ be unitary. Then (ii) (a) evidently holds. Let $Y = Z^*Z$. Then $Y \geqq 0$. Since $ZA^{-1}Z^{-1} = (ZAZ^{-1})^{-1} = B^{-1} = B^* = (ZAZ^{-1})^* = (Z^{-1})^*A^*Z^*$ it follows that $YA^{-1}Y^{-1} = A^*$.

(ii) $\Rightarrow$ (i). Let $Z \in \mathbb{F}^{nn}$, where Z is diagonal and satisfies $ZZ^* = Y$. It is easily checked that ZAZ^{-1} is unitary.

By combining Theorems 4.5 and 4.6 we obtain a corollary on which an algorithm may be based.

COROLLARY 4.7. *Let* $\mathbb{F} = \mathbb{R}$ *or* $\mathbb{F} = \mathbb{C}$, *and let* $A \in \mathbb{F}^{nn}$. *Then the following are equivalent.*

(i) *There exists a unitary matrix* $B \in \mathbb{F}^{nn}$ *such that* A *and* B *are diagonally similar.*
(ii) (a) A *is nonsingular.*
(b) $G(A^{-1}) = G(A^*)$.
(c) *If* $X = X(A^{-1} \circledcirc A^*, F, U)$, *then* $X \geqq 0$ *and* $XA^{-1}X^{-1} = A^*$.

The nonnegativity condition in (ii) (c) cannot be omitted in the above. For let $\mathbb{F} = \mathbb{R}$, and let $a, b \in \mathbb{F}$ be positive numbers with $a^2 - b^2 = 1$. Let

$$A = \begin{bmatrix} a & b \\ b & a \end{bmatrix}.$$

Then $XA^{-1}X = A^*$ where $X = \operatorname{diag}(1, -1)$, so that all other conditions in (ii) are satisfied. But every real orthogonal matrix is of the form

$$C = \begin{bmatrix} c & d \\ -d & c \end{bmatrix},$$

with $c^2 + d^2 = 1$. Let γ be the cycle $1 \to 2 \to 1$. Then $\pi_\gamma(A) > 0$ and $\pi_\gamma(C) \leqq 0$. Hence A cannot be diagonally similar to a real orthogonal matrix. The matrix A is diagonally similar to the complex orthogonal (*not* unitary) matrix

$$B = \begin{bmatrix} c & -id \\ id & c \end{bmatrix}.$$

Indeed $XAX^{-1} = B$ where $X = \operatorname{diag}(1, i)$. The results of this subsection hold for all fields $\mathbb{F}$ with involution, viz. with an automorphism $a - \bar{a}$. The elements in $\mathbb{F}$ of form $q\bar{q}$ play the role of the nonnegative elements.

4.4. Diagonal equivalence. Let $A, B \in \mathbb{F}^{nn'}$, the set of $n \times n'$ matrices. We call A *diagonally equivalent* to B if there exists a (nonsingular diagonal $X \in \mathbb{F}^{nn}$, $Y \in \mathbb{F}^{n'n'}$ for which $XAY^{-1} = B$.

For $A \in \mathbb{F}^{nn'}$ let

$$A^{+} = \begin{bmatrix} 0_{nn} & A \\ 0_{n'n} & 0_{n'n'} \end{bmatrix} \in \mathbb{F}^{n+n',n+n'},$$

where the orders of the 0 matrices are indicated by subscripts. Let $B \in \mathbb{F}^{nn'}$, $X \in \mathbb{F}^{nn}$, $Y \in \mathbb{F}^{n'n'}$, where X, Y are diagonal. As observed in [6, p. 212], $XAY^{-1} = B$ if and only if $ZA^{+}Z^{-1} = B^{+}$, where $Z = X \oplus Y$. It follows that our theorems have analogues for diagonal equivalence. The graph $G(A^{+})$ is in fact the bipartite graph of A; cf. [6]. It follows that our algorithm can easily be extended to test for the diagonal equivalence of pairs of matrices in $\mathbb{F}^{nn'}$, the simultaneous diagonal equivalence of two families of matrices, diagonal equivalence to a matrix in a subfield and diagonal equivalence to unitary matrix, see [1, Thm. 1]. Since only 0-elements are introduced in going from A to A^{+} the algorithms for diagonal equivalence are of the same complexity as those for diagonal similarity. Further theoretical details are omitted.

5. The principal algorithm. Figure 2 is a structured narrative description of an algorithm to calculate the canonical form and transformation matrix of Definitions 3.5 and 3.6. Figure 3 is a computer implementation of this algorithm in APLGOL computer language [5].

Numbers are placed on the left-hand side of corresponding steps in the two listings.

```
PROCEDURE AF←CANONICAL▵FORM A;
  ⍝ 1. INITIALIZE
    BEGIN
[1]   X(1,2,3,4,5,6,7,...,n)←1;
[2]   FOREST←1,2,3,4,5,...,n;
    END;
  ⍝ 2. TRAVERSE FOREST
[3]  WHILE FOREST IS NONEMPTY DO
       BEGIN
[4,5]    Remove an element from FOREST and define TREE
           to be a list whose only entry is this element;
  ⍝ 3. TRAVERSE TREE
         REPEAT
[6,7]       Remove an element from TREE and set BRANCH
             equal to this element;
[8]         Search the row indexed by BRANCH for
             nonzero elements whose column index is in FOREST.
             Set BRANCHES equal to this index set;
[9]         IF BRANCHES IS NONEMPTY THEN
                BEGIN
[10]              X(BRANCHES)←A(BRANCH;BRANCHES)×X(BRANCH);
[11,12]           Remove indices in BRANCHES from
                    FOREST and place in TREE;
                END;
[8']        Search the column indexed by BRANCH for
              nonzero elements whose row index is in FOREST
              and set Branches equal to this index set;
[9']        IF BRANCHES IS NONEMPTY THEN
                BEGIN
[10']             X(BRANCHES)←X(BRANCH)÷A(BRANCHES;BRANCH);
[11',12']         Remove indices in BRANCHES from
                    FOREST and place in TREE;
                END;
[13]      UNTIL TREE IS EMPTY;
       END;
[14]  Print the diagonal of the transforming matrix X(1,2,3,.....,n);
[15]  Take the Hadamard product of A with the outer product of
                                  -1          -1
       X(1)...X(n) and X(1).....X(n)  to form AF;
[16]  Print the canonical form AF;
END PROCEDURE
```

FIG. 2

```
      ∇PROCEDURE AF←CANONICAL∆FORM A,FOREST,TREE,BRANCH;
[1]     X←(1↑⍴A)⍴1;
[2]     FOREST←⍳⍴X;
[3]     WHILE (⍴FOREST)≠0 DO
          BEGIN
[4]         TREE←1↑FOREST;
[5]         FOREST←1↓FOREST;
            REPEAT
                BRANCH←1↑TREE;
[6]             TREE←1↓TREE;
[7]             1 TRAVERSE,A[BRANCH;];
                  1 TRAVERSE,A[;BRANCH];
[13]        UNTIL (⍴TREE)=0;
          END;
[14]    □←X;
[15,16] □←AF←A×X∘.=X;
      END PROCEDURE
      ∇PROCEDURE E TRAVERSE VECTOR,BRANCHES;
[8,8']  BRANCHES←(VECTOR[FOREST]≠0)/FOREST;
[9,9']  IF (⍴BRANCHES)≠0 THEN
          BEGIN
[10,10']    X[BRANCHES]←X[BRANCH]×VECTOR[BRANCHES]*E;
[11,11']    TREE←TREE,BRANCHES;
[12,12']    FOREST←(~FOREST∈BRANCHES)/FOREST;
          END;
      END PROCEDURE
```

FIG. 3

Computational complexity. If A is a $n \times n$ matrix such that $G(A)$ has t components then the execution of this algorithm results in $6n-2t$ storage operations, $n-t$ multiplications or divisions, and fewer then $2n+t+n^2$ but more then $4n+t-1$ logical operations. Table 1 provides a statement by statement accounting of the complexity.

Steps 1 and 2 are not included in this accounting since the vectors X and FOREST can be initialized prior to execution.

Logical operations are simplified by avoiding the concepts used in analyzing directed graphs. The algorithm involves only straightforward pointer maintenance. Backtracking and recursive executions is avoided. In addition this algorithm does not require precomputation of the column numbers of the nonzeros in each row as is the case in many algorithms in combinatorial matrix theory, e.g., the Duff–Reid implementation of Tarjan's algorithm for the block triangulization of a matrix [2].

TABLE 1

Statement number	Number of storage operations	Number of multiplications or divisions	Number of logical operations
(3)			t
(4)	t		
(5)	t		
(6)	n		
(7)	n		
(13)			n
(9, 9′)			$2n$
(10, 10′)	$n-t$	$n-t$	
(11, 11′)	$n-t$		
(12, 12′)	$n-t$		
(8, 8′)	$n-t$		$n-1 \leqq$ logical op $\leqq nXn-n$
Total	$6n-2t$	$n-t$	$4n+t-1 \leqq$ logical op $\leqq 2n+t+n^2$

6. Applications. We have applied the algorithm of § 5 for finding the canonical form under diagonal similarity to yield the tests shown in Table 2.

TABLE 2

Test	Justification
Diagonal similarity of a pair of matrices	Corollary 3.11
Simultaneous diagonal similarity of a family of matrices	Theorem 4.3
Diagonal similarity of a real matrix to an orthogonal matrix	Corollary 4.7
Extensions of the three algorithms above to the corresponding algorithms for diagonal equivalence	§ 4.4

The first of these algorithms is described in Fig. 4 in APLGOL notation. Detailed descriptions of some other algorithms are contained in the authors' technical report.

```
PROCEDURE A DIAGONAL∆SIMILARITY∆TEST B,HF;
  ⍝VERIFY G(A)= G(B)
  IF ∧/,(A≠0)=(B≠0) THEN
    BEGIN
      ⍝ COMPUTE THE HADAMARD QUOTIENT FOR A AND B.
      H←A÷B+B=0;
      ⍝ COMPUTE THE CANONICAL FORM HF
      HF←CANONICAL∆FORM H;
      ⍝ VERIFY ALL THE ENTRIES OF HF ARE EITHER ZERO OR ONE
      IF ∧/,HF∈0,1 THEN
        BEGIN
          □←'MATRICES ARE DIAGONALLY SIMILAR';
          ⍝ PRINT OUT THE DIAGONAL OF THE SIMULARITY TRANSFORMATION
          □←X;
        END
      ELSE
        □←'MATRICES ARE NOT DIAGONALLY SIMILAR';
    END
  ELSE
    □←'THE GRAPHS OF THE MATRICES ARE UNEQUAL';
END PROCEDURE
```

FIG. 4

REFERENCES

[1] A. BERMAN, B. N. PARLETT AND R. J. PLEMMONS, *Diagonal scaling to an orthogonal matrix*, this Journal, 2 (1981), pp. 57–65.

[2] I. S. DUFF AND J. K. REID, *An implementation of Tarjan's algorithm for the block triangularization of a matrix*, ACM Trans. Math. Software, 4 (1978), pp. 137–147.

[3] G. M. ENGEL AND H. SCHNEIDER, *Cyclic and diagonal products on a matrix*, Linear Alg. Appl., 7 (1973), pp. 301–335.

[4] M. FIEDLER AND V. PTÁK, *Cyclic products and an inequality for determinants*, Czechoslovak Math. J., 19 (1969), pp. 428–450.

[5] R. KELLY AND S. J. WALTERS, APLGOL-2, *A structured programming system for* APL, IBM Palo Alto Sci. Rep. G 320-3318, 1973.

[6] D. B. SAUNDERS AND H. SCHNEIDER, *Flows on graphs applied to diagonal similarity and diagonal equivalence for matrices*, Discrete Math., 24 (1978), pp. 205–220.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0005 $01.00/0

NEW METHODS FOR EVALUATING DISTRIBUTION AUTOMATION AND CONTROL (DAC) SYSTEMS BENEFITS*

JOHN PESCHON† AND DALE ROSS†

Abstract. The decade ahead will be one of heightened concern for costs versus benefits to end users of electric energy. More than in the past, the distribution planner will be concerned about investment costs, operating efficiency and reliability of service. The advent of new dispersed generation, storage and control technologies for distribution systems will change not only the alternatives available to the planner, but also the planning methods themselves.

This paper summarizes the development of new distribution planning methods. In particular, methods have been developed for both expansion planning and operations planning of radial distribution systems. A particular application of distribution automation and control will be for temporary distribution system reconfiguration during either forced outages or maintenance/construction-related outages. Remotely controlled switches can be used to transfer load among radial feeders during construction, maintenance or other service interruptions—thereby reducing or preventing outages for many customers. This paper describes computerized methods for evaluating the reliability benefits of such advanced distribution systems.

1. Introduction. Distribution reliability has traditionally been a major concern of power system planning and operations. An EPRI report [1] suggests that reliability considerations should be included in all distribution system planning due to the fact that 98% of all customer interruptions are caused by trouble on distribution systems. Recently CEA recognized the importance of this topic with the publication of an engineering guide [2] on distribution reliability. A summary of present applications of reliability evaluation in distribution systems was given at an IEEE panel session on Distribution System Reliability [3], [4], [5].

The increased interest in distribution reliability is, in part, associated with the increasing proportion of distribution systems that are operating at considerably higher voltages than was the practice a decade or two ago. The move to a higher primary voltage has been reported [3], [4] to increase the potential for more widespread customer interruptions due to longer feeder lengths. In order to realize the advantages of higher distribution voltages without degradation in reliability due to longer feeder lengths, it is necessary to apply new approaches. One approach is to use automatic sectionalizing under supervisory control, with the capability to close ties to other feeders—thereby minimizing both the duration and extent of outages [5]. Some utilities [4], [6] have established design criteria in which load area feed reliability requirements are a function of load density. At low densities only single feed may be required; at higher densities, dual feed may be required—with a normally open tie switch to a second source; at even higher densities, automatic changeover to a second feed may be required.

This paper discusses a recently developed [7] computerized model which can aid planners in applying new approaches to achieving reliable electric power distribution systems. The model, called SWITCH, optimizes the reconfiguration of radial feeder systems during either emergency or planned outages. The SWITCH model can be used by electric utilities for a variety of planning and operating problems, including:

- Performing cost/benefit studies of proposed schemes for sectionalizing and reconfiguration of radial distribution systems.

* Received by the editors June 18, 1981. Portions of this paper were presented at the 1980 Reliability Conference for the Electric Power Industry, Madison, WI, May 1980, and at the 140th meeting of the Edison Electric Institute Transmission and Distribution Committee, Oklahoma City, October 1979.

† Systems Control, Inc., 1801 Page Mill Road, Palo Alto, California 94304.

- Developing operating plans for the use of switches to reconfigure feeders during planned outages (such as for maintenance or construction).
- Performing comparative evaluations of the reliability of service provided in different portions of a utility's service area to different classes of customers.
- Planning remote control systems and other future distribution automation functions.

The capabilities of the SWITCH model have been demonstrated in a variety of case studies. Some of the case study results point very positively to its usefulness:

- Several SWITCH case studies indicated that automatic switching has the potential to reduce customer-hours of interruption by 20% to 90% depending on the nature of the initiating fault.
- The computational efficiency of the model is quite good—complex contingencies can typically be evaluated using less than 30 sec. of computer time.

2. Overview of the evaluating method (the SWITCH model). Given a geographic area served by one or more substations having many radiating feeders, the planner is interested in determining the merits of different sectionalizing and switching schemes. A general procedure for doing this is depicted in Fig. 1. The procedure systematically examines the effects of circuit reconfiguration to each of a set of selected component outage contingencies. The set of contingencies can be either exhaustive or a statistical sample from among all contingencies. By running through this procedure twice, one can evaluate the reduction in outage consequences attributable to the switching capability.

The key element of the general procedure is an optimization program. This program, called SWITCH, determines how best to use the available switches to minimize outage consequences. The objective of SWITCH is to obtain a *sequence* of reconfigurations that minimize the number of unserviced customer hours/unserviced energy during the period when a fault (contingency) is being isolated and repaired. The energized network must satisfy:

- demands of connected loads,
- current capacities,
- voltage limits.

A *sequence* of reconfigurations is found because the following general seven-step scenario is assumed to apply to each fault occurrence. These seven steps are simulated via the SWITCH procedure.

Step 0. The first protective device (e.g., sectionalizer, recloser) above[1] the fault trips, deenergizing all loads below the device.

Step 1. The fault is isolated using remotely controlled switches as follows: (a) open remote switches below[1] the fault; (b) if protective device can be remotely controlled, then further isolate the fault by opening the first remotely-controlled switch above the load and closing the protective device. Loads between the protective device and remotely controlled switch above the fault are not serviced.

Step 2. Loads that are both below the isolating switches of Step 1 and below the fault can potentially be reconnected at this time using remotely-controlled switches. The reconfiguration optimization algorithm (to be discussed below) is used to determine the load to reconnect so as to satisfy the current capacity and voltage limit constraints for the entire network. This

[1] "below" means further from the source, while "above" means closer to the source.

is the "Stage 1" reconfiguration of three reconfigurations that occur.

Step 3. The fault is isolated further using both manually and remotely-controlled switches. Manually-controlled switches immediately above the fault and below the fault are opened.

Step 4. Unserviced load-points can potentially be reconnected at this time using remotely-controlled switches. The reconfiguration optimization algorithm is used to determine which load-points can be reconnected and still satisfy the system constraints. This is the "Stage 2" reconfiguration.

Step 5. Both remotely and manually-controlled switches are now employed to reconnect load using the reconfiguration optimization algorithm. This is the "Stage 3" reconfiguration.

Step 6. Following the repair of the fault, the network is returned to its initial configuration.

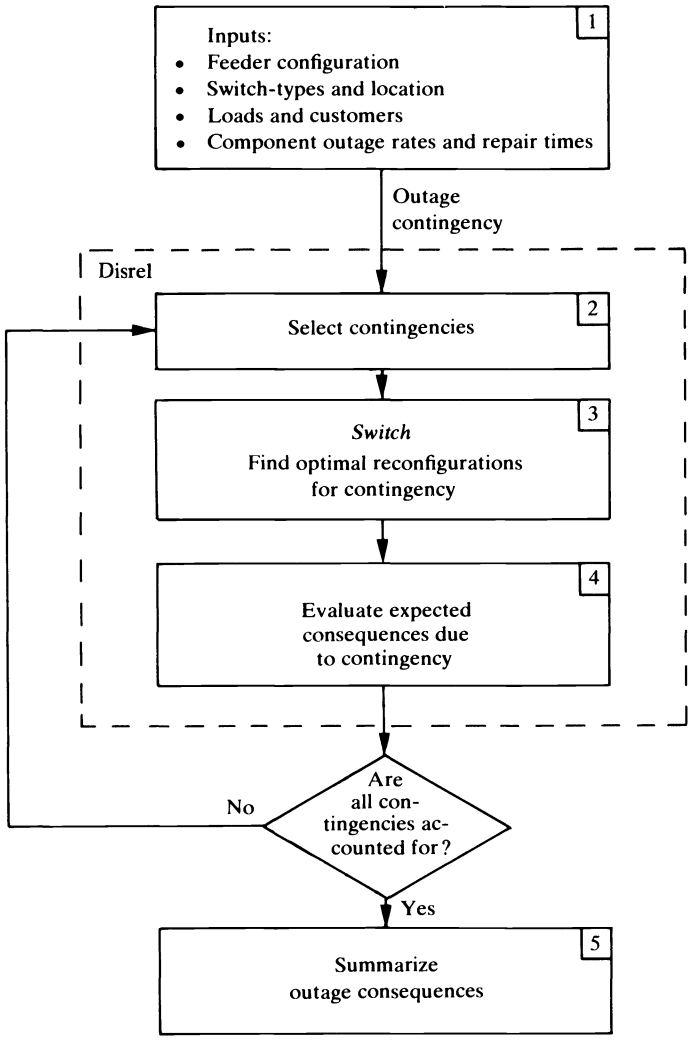


FIG. 1. *General procedure for evaluating reliability benefits of manual or automatic switching.*

Figure 2 depicts how the number of customers on outage, for example, diminishes with each step of the SWITCH procedure. The times T_i taken for each step are a function of the speed with which the manually- and remotely-controlled switching operations can be performed. In Fig. 2, the area between the "staircase" and the "without reconfiguration capability" characteristic is the customer-hours saved by the switching operations. If this and/or the corresponding savings in unserviced energy are monetized—the SWITCH procedure is seen to be the basis of cost/benefit method. The reliability benefits of manual and automatic switches can be compared with their costs. If this is done for all selected contingencies, per Fig. 1, the total net benefit for the distribution system can be estimated.

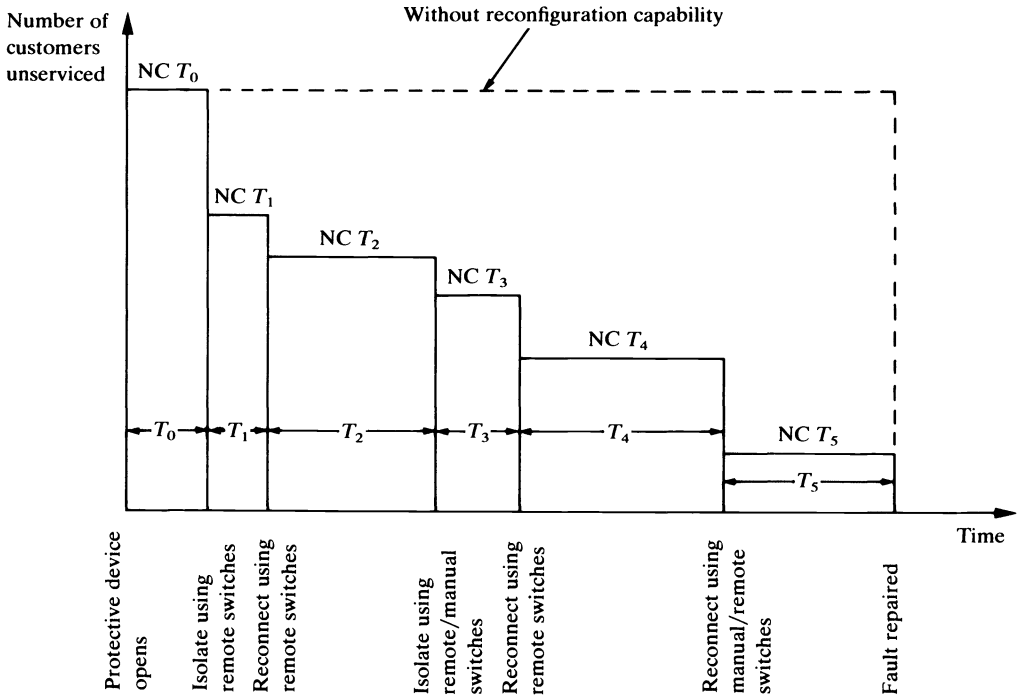


FIG. 2. *Example of number of customers unserviced during switch procedure.*

3. Detailed logic within the SWITCH model. The benefits of advanced distribution control systems stem from their capability once a power outage has occurred to reservice as many utility customers as is feasible as quickly as possible. Decisions must therefore be made regarding:

- which unserviced loads to transfer to energized feeders,
- which unserviced loads to leave unserviced if it is not feasible to reservice all unserviced loads.

The techniques for making these decisions are outlined below.

The reconfiguration optimization algorithm. As described previously, emergency reconfiguration to a fault is assumed to take place in three stages. The basic optimization problem to be solved in each stage is that of using the system's switches (viz., those available in that particular stage) to perform a radial reconfiguration that minimizes a weighted sum of the unserviced cutsomer-hours and the unserved energy and that satisfies the emergency voltage and current constraints.

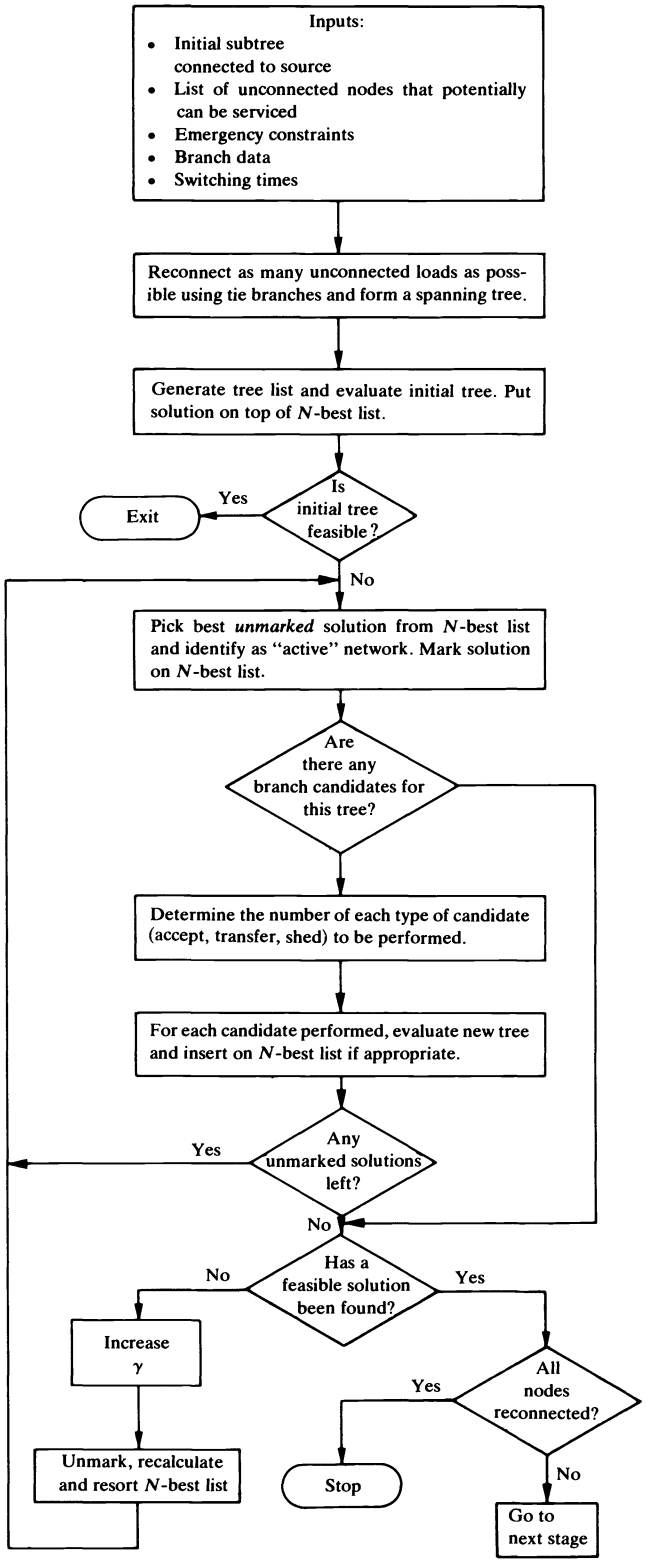


FIG. 3. *The reconfiguration optimization algorithm.*

The algorithm that is used to perform the reconfiguration optimization (for a given stage) is depicted in Fig. 3. The material below explains the logic depicted in the figure.

The optimization algorithm starts with an initial radial reconfiguration that reconnects as many of the unenergized nodes as possible without regard to the constraints on voltage or current. Before creating additional reconfigurations, the algorithm computes an objective function value OBJ for the initial reconfiguration. That is:

$$\text{OBJ} = \gamma \ \text{VALINF} + \text{VALSW}, \tag{1}$$

where VALINF is the value of the "infeasibility" of the initial reconfiguration and VALSW is the reconfiguration value (i.e., a measure of the unserved energy/customer hours resulting from this reconfiguration). VALINF is given by:

$$\text{VALINF} = \max_{j \in B} (0, -(\bar{I}_j - I_j)) + \max_{k \in N} (0, -(V_k - \underline{V})), \tag{2}$$

where B is the set of all energized feeder branches in the reconfiguration and N is the set of all served nodes. VALSW includes the terms in the unserviced energy/customer-hours affected by the reconnection at this stage. Now the *total* (i.e., including all stages) objective to SWITCH is to minimize

$$\sum_{i=0}^{5} [\alpha \text{NC}_i + \beta \text{UP}_i] T_i, \tag{3}$$

where

NC_i = number of customers without service during step i,
UP_i = unserviced power during step i,
T_i = time it takes to reconfigure for step i;

α and β are constants reflecting the importance the planner places on unserviced customers-hours and unserviced energy, respectively. At a particular reconfiguration stage, only two of the terms in (3) are optimized. For example, at Stage 1 where we reconnect using remote switches following remote isolation (see Fig. 2) only the terms with $i = 1$ and 2 are affected by the reconfiguration. For this case

$$\text{VALSW} = [\alpha \text{NC}_1 + \beta \text{UP}_1] T_1 + [\alpha \text{NC}_2 + \beta \text{UP}_2] T_2. \tag{4}$$

Similarly, VALSW for Stage 2 contains terms with $i = 3$ and $i = 4$, and for Stage 3 it contains terms with $i = 4$ and $i = 5$.

The constant γ in the OBJ function is used to ensure that a feasible solution is obtained. The larger γ, the more the algorithm will concentrate upon eliminating voltage and current constraint violations.

The value OBJ is assigned to the initial configuration, and it becomes the first entry on an "N-best list" of reconfigurations. The N-best list is a list of the N-best solutions found so far, where N is an input parameter to the algorithm.

Candidate branches. Once the starting conditions have thus been established, the algorithm proceeds to alter the network by considering the energizing/deenergizing of feeder branches on three "candidate branch" lists. (The total number of candidates to be considered is a user input.) The three types of candidate branches are:

Load-transfer candidate. This is a branch which when *energized* will connect two nodes that are already in the serviced network. It would be advantageous to energize this type of branch if doing so enlarges the set of nodes and branches for which voltage

and current constraints are satisfied. (Of course, whenever such a branch is energized, a loop is formed and some branch in the loop must be deenergized in order to maintain a *radial* configuration.)

Load-accept candidate. This is a branch which when *energized* will connect a node in the already serviced network with one in the set of unserviced nodes. Obviously, if this connection retains voltage and current feasibility, then it enlarges the set of nodes that can be serviced.

Load-shed candidate. This is a branch which when *deenergized* separates two nodes in the already serviced network—leaving one of them without service. It may be necessary to disconnect some nodes in order to serve others with feasible voltages and current.

In order to decide which candidate branches are the better choices, it is necessary to establish some measure of the advantage obtained by energizing (or deenergizing) each one. This measure is called the *value* of the branch. The value of a branch is in turn computed from voltage and current margins, which are defined below.

Current and voltage margins. The most important variables for identifying candidate branches are current and voltage margins. The *current margin* at a node k is the amount of additional load that can be added to the node without violating the current constraints anywhere in the radial system. If there originally is a current constraint violation in the path from the node k back to the source, then the current margin will be negative and it will be equal to the amount of load that must be subtracted from the node k so that the current constraints are satisfied in the path back to the source. Mathematically, the current margin at node k is given by

$$\mathrm{MC}_k = \min_{j \in S_k} \{\bar{I}_j - I_j\},$$

where S_k is the set of nodes in the path from k to the source; I_j is the current in branch $(a(j), j)$, $(a(j)$ is the predecessor of node j in the system or equivalently the node above $j)$ and $\bar{I}_j$ is the maximum current allowed in branch $(a(j), j)$. The current margin for all nodes can be calculated recursively, starting at the source, by

$$\mathrm{MC}_j = \min \{\mathrm{MC}_{a(j)}, \bar{I}_j - I_j\}.$$

The *voltage margin* is similarly defined. It is the amount of additional load that can be added to a node without violating any of the voltage constraints. If there is originally a voltage constraint violation, then the voltage margin is negative and is equal to the amount of load that must be removed from the node in order to satisfy the voltage constraints. While the current margin at node k only depends on the currents and current constraints in the branches in that path from node k back to the source, the voltage margins are affected, in general, by the voltages and voltage constraints in all nodes in the same limb as node k. Mathematically, the voltage margin is given by

$$\mathrm{MV}_k = \min_{m \in L_k} \{(V_m - \underline{V})/\rho_{mk}\},$$

where L_k is the set of nodes in the limb containing k, V_m is the voltage at node m, $\underline{V}$ is the voltage constraint and ρ_{mk} is the impedance of the path common to S_k and S_m (S_k denotes the path from node k back to the source). A recursive algorithm to calculate voltage margin is described in [7].

The purpose of margins in the method is two-fold:

- to specify how much additional load a node can accept without violating a constraint,

- to specify how much load must be shifted to a different part of the network to bring the network into feasibility.

Current and voltage margins can be used for both of these; however, voltage margin is not very helpful for the second purpose. The problem with using voltage margin to identify loads to transfer is that voltage margin will find the load to be subtracted from a node to satisfy *all* the voltage constraints on a feeder. If there are two voltage violations and the paths back to the source from the nodes violating these constraints barely intersect, then the voltage margins obtained carry little information.

To obtain information for performing load transfers, a new margin concept is introduced—the *worst-node voltage margin*. To define this, we first determine for each limb a lower bound on the amount of current that must be transferred to cause satisfaction of the voltage constraints, and we identify the node j^* at which this occurs. Let

$$\mathrm{MV}_l^* = \min_{m \in N_l} \{(V_m - \underline{V})/\rho_m\}, \qquad j_l^* = \arg\min_{m \in N_l} \{(V_m - \underline{V})/\rho_m\},$$

where N_l is the set of nodes on feeder l and where ρ_m is the impedance from node m to the source. Since it is not always possible to transfer load at the node where a violation is taking place, we are interested in the amount of load that must be transferred from other nodes to cause satisfaction of the voltage constraint at this "worst node" j_l^*. We therefore define the worst node voltage margin as the amount of load in amps that must be removed from a node to cause satisfaction of the voltage constraint at j_l^* or

$$\overline{\mathrm{MV}}_k = (V_{jl}^* - \underline{V})/\rho_{j_l^* k},$$

where l is the feeder containing k and ρ_{jlk}^* is the impedance of the path to the source that is in common to nodes j_l^* and k.

Other variables used in determining if a load transfer will improve feasibility are the voltage at a node and the current flowing into a node. These quantities are useful in determining if energizing a branch will result in a constraint violation at other points in the network. The manner in which the various criteria are used is discussed below.

Value of a load-transfer candidate. The value of a load-transfer candidate branch (k, l) is a measure of the degree to which energizing that branch will:

- improve the current margin at k,
- not cause the current-carrying capacity of the branch to be exceeded,
- improve the voltage at k and
- improve the worst-node voltage margin at k.

Functionally, the value VLT_{kl} is specified as:

$$\begin{aligned}\mathrm{VLT}_{kl} = {} & a_{\mathrm{MC}}(\mathrm{MC}_l - I_k, \mathrm{MC}_k) + a_C(\bar{I}_{kl} - I_k) \\ & + a_V(V_l - I_k(R_{kl} + \rho_l) - \underline{V}) + a_{\mathrm{MV}}(\overline{\mathrm{MV}}_l, \overline{\mathrm{MV}}_k),\end{aligned} \tag{5}$$

where MC_k and $\overline{\mathrm{MV}}_k$ are the previously defined margins, I_k is the total current flowing toward node k from the source direction, V_l is the voltage at node l, $\bar{I}_{kl}$ is the current-carrying capacity of branch (k, l), R_{kl} is the impedance of the entire path

from node l back to the source feeding that node. The functions $a_{MC}(\cdot)$, $a_C(\cdot)$, $a_V(\cdot)$, $a_{MV}(\cdot)$ are used as follows (see [7] for details):

- $a_{MC}(\cdot)$ establishes a value for the degree of current margin improvement.
- $a_C(\cdot)$ establishes a value for the acceptability of the current loading on branch (k, l).
- $a_V(\cdot)$ establishes a value for the acceptability of the (estimated) voltage at node l, once load has been transferred there from node k.
- $a_{MV}(\cdot)$ establishes a value for the degree of worst-node voltage margin improvement.

Value of a load-accept candidate. The value of a load-accept candidate branch (k, l) is a measure of the degree to which energizing that branch will:

- provide an acceptable (i.e., positive) current margin at node l,
- provide an acceptable (i.e., positive) voltage margin at node l,
- load branch (k, l) safely within its current-carrying capacity and
- provide an acceptable voltage at node k.

Functionally, the value is specified as:

$$\mathrm{VLA}_{kl} = a_C(\bar{I}_{kl} - I_k) + a_V(V_l - I_k(R_{kl} - \rho_l) - \underline{V}),$$

where the functions $a_C(\cdot)$ and $a_V(\cdot)$ are those discussed previously.

Value of a load-shed candidate. The value of a load-shed candidate branch (k, l) is a measure of the degree to which deenergizing that branch will:

- eliminate a poor current margin at node k,
- eliminate a worst-node voltage margin at node k.

Functionally, the value is specified at:

$$\mathrm{VLS}_{kl} = e_C(\mathrm{MC}_k) + e_V(\overline{\mathrm{MV}}_k, \mathrm{MV}^*_{L_k}),$$

where the function $e_C(\cdot)$ establishes a value for the elimination of a negative current margin at node k, and the function $e_V(\cdot)$ establishes a value for the elimination of a negative worst-node voltage margin at k. These functions are defined in [7]. MC_k and $\overline{MV}_k$ are the previously defined current margin and worst-node voltage margin. $MV^*_{L_k}$ is the voltage margin at the worst node within the feeder that contains node k.

Finding the best reconfiguration. Once all candidate branches have been identified (that is, have been classified as either a load-transfer, load-accept or load-shed type and have had their values computed), then ranked lists of each type are prepared. These ranked lists then become the starting inputs to an algorithm that forms successive emergency reconfigurations, examining each in sequence until convergence to a best configuration has been obtained.

4. Sample results. The SWITCH evaluation procedure has been tested in a number of case studies; one of these is summarized below (other cases are in [7]). The study was of a large distribution system serving a 13 square mile area through a substation with a load of 58 MW (the area therefore has an average load density of 4.5 MW/square mile). The substation serves 12,400 customers with an 80% residential load in an urban area; the area also has some small industrial and commercial loads and a large shopping mall.

One contingency that was selected for study was a fault on branch (CH01, CH02) in Fig. 4. Assuming that automatic fault location equipment is installed in the system, 2.0 minutes time is taken for locating the fault, after which the remotely controlled breaker is left open on branch (CH01, Feeder 2), causing nodes CH01, CH02, CH22,

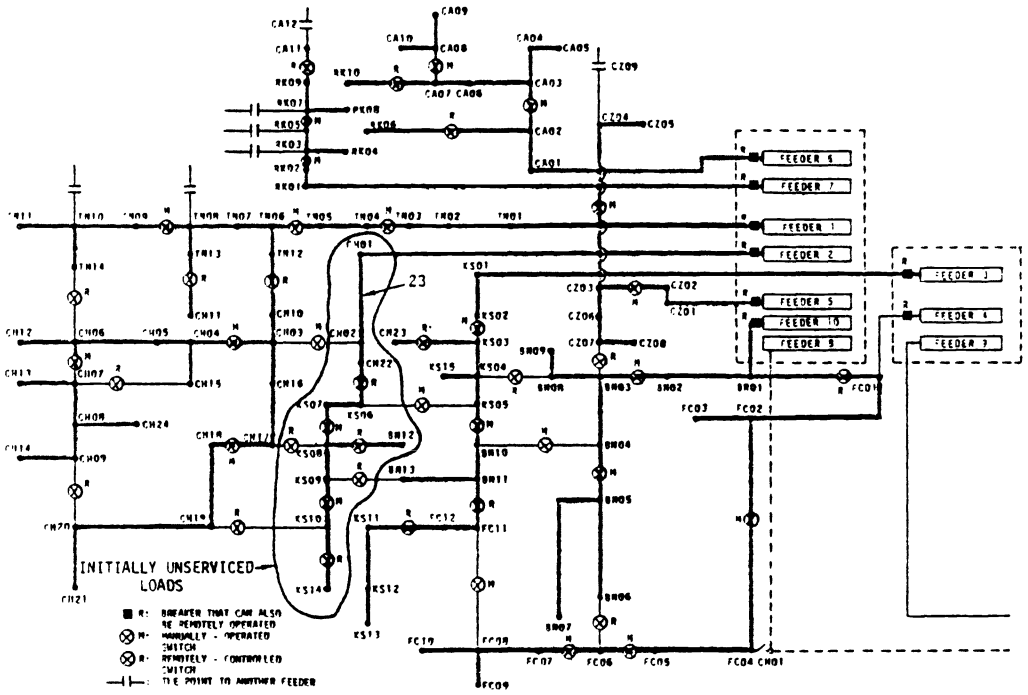


FIG. 4. *The large distribution system with fault on arc* 23.

KS06, KS07, KS08, KS09, KS10, KS14 and BN12 to be unserviced—resulting in 1651 unserviced customers and 128 kW unserviced load. Because the fault cannot be further isolated, Stage 1 reconnection of load via remotely-controlled switches reconnects as many of the unserviced nodes as possible. In this case, nodes KS06, KS07, KS08, KS09, KS10, KS14 and BN12 can be reconnected by using the remote switching capability on branch (CH22, KS06) and branch (KS09, BN13). In this stage, node CH01 remains unserviced due to the open breaker; nodes CH02 and CH22 remain unserviced due to the fault and the lack of switching capability above node CH22. Stage 1 has been completed and Fig. 5 shows the resulting configuration. The net result obtained was that all but three nodes were reconnected using two remote switches in 1.5 minutes thereby reducing the unserviced customers by 1491 (from 1651 to 160) and the unserviced load by approximately 100 kW (from 128 to 28 kW).

During Stage 2 (further isolation of the fault via manually operated switches) node CH01 was reconnected after 157 minutes by manually isolating the faulted branch (CH01, CH02) therefore reducing the unserviced customers by 80 (from 160 to 80) and the unserviced load by 6 kW (from 28 to 22 kW). Only nodes CH02 and CH22 are left unserviced. These two nodes must be reconnected together because there is no switching capability above node CH22 to allow them to be separated. The SWITCH algorithm tries to reconnect them to node KS06 but the resulting configuration is infeasible. Stage 2 has been completed and Fig. 6 shows the resulting configuration. The net result obtained by SWITCH was that node CH01 was reconnected during the remote and manual isolation step. No new switching was performed during the remote reconnection step.

During Stage 3 (remote plus manual reconnection) again nodes CH02 and CH22 are together reconnected to node CH03 but the resulting configuration is infeasible.

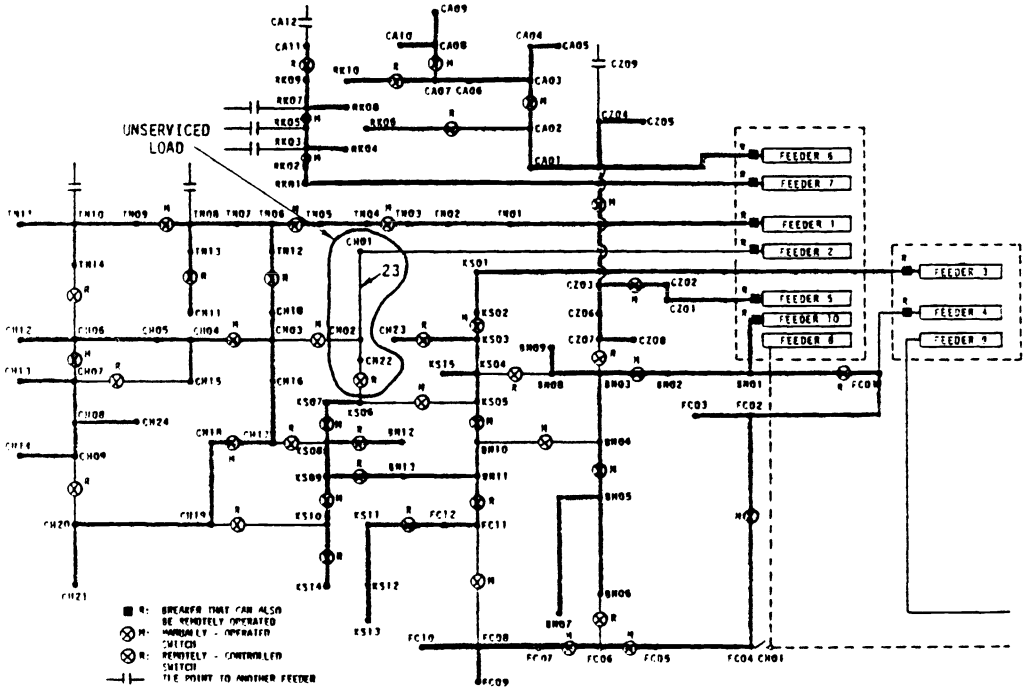


FIG. 5. *The large distribution system with fault on arc* 23; *after stage* 1.

Since there are no other reconnection possibilities, Stage 3 is completed with no changes to best configuration found in Stage 2. A configuration with all nodes serviced cannot be found given the present switching capability. Therefore, 80 customers and 22 kW load are left unserviced for 180 minutes (the assumed fault repair time).

Table 1 gives a summary of the number and type of switches operated, the switching time, the number of unserviced customers and the amount of unserviced load for each reconfiguration stage for this case.

From the data in Table 1, it can be shown (assuming fault repair is conducted simultaneously with the reconfiguration) that the reconfiguration saves 4,413 customer-hours relative to leaving all customers on outage until the fault is repaired. That is, for each occurrence of this fault, there would be 4,953 unserviced customer hours if no reconfiguration capability existed but only 540 customer hours with reconfiguration. Eighty-nine percent (89%) of the outage time is saved.

Other cases were also examined. For instance, reconfiguration was found to save 52% and 21% of the unserviced customer-hours for faults on branches (FC04, FC05) and (KS07, KS08), respectively. The magnitudes of these savings were 8,284 and 1,052 customer-hours, respectively.

The computer cpu time (on a UNIVAC 1108) required for the study of the tree faults was very modest—ranging from 3 to 37 seconds for each fault.

5. Conclusions. A computerized method, the SWITCH procedure, has been developed for evaluating the reliability benefits of sectionalizing and reconfiguration capability in radial distribution systems. SWITCH has a number of uses. Our case studies focused on only one of these—single contingency analysis in systems having remote and manual switching capability. Space limitations did not permit us to illustrate

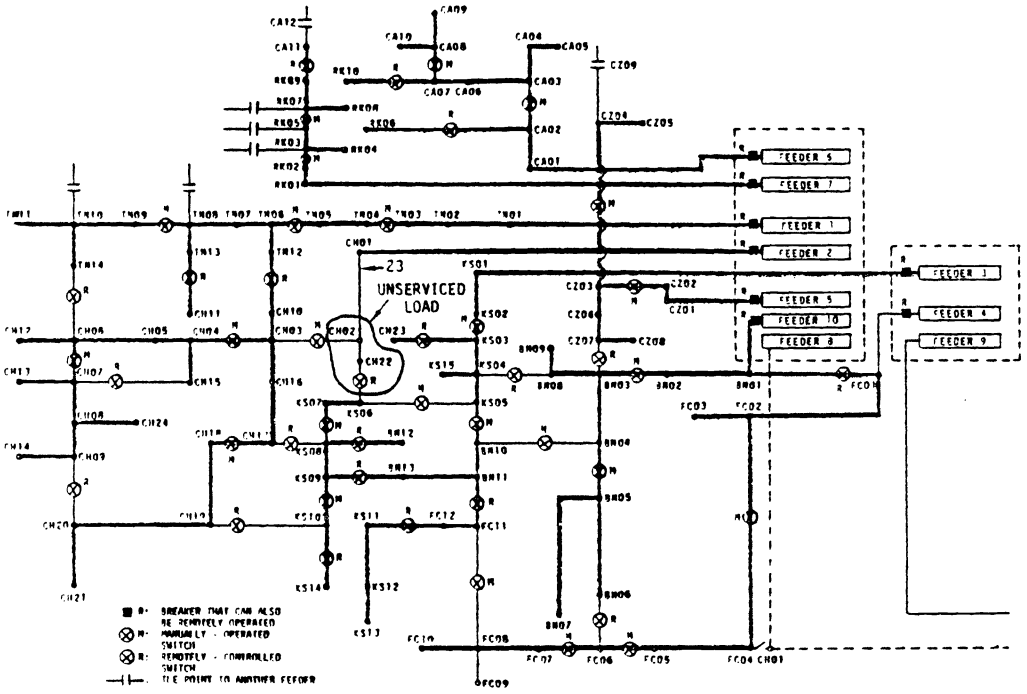


FIG. 6. *The large distribution system with fault on arc* 23; *after stage* 2.

TABLE 1

Reconfigurations performed at each stage for the large distribution system with fault on arc 23.

	No. remote switches (after)	No. manual switches (after)	Duration (min)	Unserviced customers (before)	Unserviced load (kW) (before)	Unserviced customers (after)	Unserviced load (kW) (after)
STAGE 1							
Remote fault location/ isolation	—	—	2.0	1651	127.67	1651	127.67
Remote reconnection	2	0	1.5	1651	127.67	160	28.41
STAGE 2							
Remote/manual fault isolation	—	—	156.3	160	23.41	80	21.32
Remote reconnection	0	0	0.0	80	21.52	80	21.52
STAGE 3							
Remote/manual reconnection	0	0	0.0	80	21.52	80	21.52
Repair fault	—	—	180.0	80	21.52	0	0.0

other potential applicatios of the SWITCH model. However, the two tables below provide a summary of the various uses to which SWITCH could be placed. Specifically,

Table 2 summarizes potential applications of SWITCH to operating problems, and Table 3 summarizes its potential applications to planning problems.

Further details on this work presented in this paper are contained in [7].

TABLE 2
Potential applications of the switch model to the operation of electric power distribution systems

Potential Application	Description of SWITCH use
Operation of future distribution automation and control (DAC) systems	The SWITCH logic is a prototype that could become software for real-time control of switches in distribution systems. Such automation may give reliability benefits (from faster response to faults).
System reconfiguration during planned outages	SWITCH could be used to establish operating plans for the use of switches to reconfigure feeders during planned outages (such as for maintenance or construction).
Load shedding and restoration	In emergencies, distribution automation and control (DAC) could be used to drop large blocks of load. During restoration, the SWITCH type of logic might be used in a DAC system to restore (via switching operations) service to small blocks of load as called for in the restoration plan.

TABLE 3
Potential applications of the switch model to the planning of electric power distribution systems.

Potential application	Description of SWITCH use
Substation emergency capacity rating	Many utilities use a two-transformer substation configuration. Traditionally, maximum emergency load (emergency capacity rating) considered feasible for such a substation is limited by the 24-hour rating of each transformer. But, by using the SWITCH model, one can study the transformer outage contingencies and determine how much *additional load* can be serviced by using tie switches to adjacent feeders and by using sectionalizing switches for emergency reconfiguration. This additional load when added to the single-transformer capacity yields a less conservative—and more accurate—emergency capacity rating for the substation.
Cost/benefit analysis of distribution automation and control (DAC)	If the planner can at least establish a range for the monetary value of savings in customer-hours of outage or unserved energy—then the SWITCH model can be used to estimate the benefits of different sectionalizing and switching schemes. One would use SWITCH to systematically examine the effects of circuit reconfiguration to each of a set of component outage contingencies. The set of contingencies can be either exhaustive, or a statistical sample from among all contingencies. By comparing the customer-hours of outage and unserviced energy for the set of contingencies with and without the DAC, one can evaluate the benefits of the DAC. These benefits can then be compared to the DAC costs.
Cost/benefit analysis of installation of sectionalizing and tie switches	Similar to the DAC benefit/cost analysis—except that all switches are assumed to be manually operated.
Reliability analysis	The effect of switching capability in the distribution system upon the duration of customer outages can be studied using SWITCH.

Acknowledgment. The individual assistance and guidance of Tobey A. Trygar, U.S. DOE Project Manager, is gratefully acknowledged. Our colleagues at Systems Europe, S.A., also contributed significantly to this work; in particular, we have benefited from the assistance of Claude Dechamps, Rik Nuytten and Jacques Vankelecom.

REFERENCES

[1] *Analysis of distribution R & D planning*, EPRI RP 329, Systems Control, Inc., Palo Alto, CA, 1975.

[2] *Distribution System Reliability Engineering Guide*, Distribution System Reliability Engineering Committee, Canadian Electric Association (CEA), March 1976.

[3] P. GHOSE, *Distribution system reliability*, IEEE PES Winter Power Meeting, New York, January 31, 1978.

[4] R. J. BUTRYM, *Distribution reliability considerations at Wisconsin Electric*, ibid.

[5] J. L. KOEPFINGER, *Automation and reliability—H. V. distribution circuits*, ibid.

[6] A. R. PEARSON AND P. B. JONES, *Technical and economic factors affecting the evaluation of reliability of supply to load groups*, Trans. IEEE Conference on Reliability of Power Supply Systems, IEEE Conference Publication, 148, February 1977, pp. 62–65.

[7] D. W. ROSS, et al., *Development of advanced methods for planning electric energy distribution systems*, Draft Final Report, Systems Control, Inc., Palo Alto, California, November 1979 (prepared for U.S. Department of Energy under contract ET-78-C-03-1845).

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0006 $01.00/0

PRACTICAL APPLICATIONS OF DISCRETE MATHEMATICAL PROGRAMMING IN EXXON*

WILLIAM P. DREWS†

Abstract. Applications of discrete mathematical programming may be subdivided into those involving economies of scale, those involving mutually exclusive variables and those involving nonconvexity in the constraint set. Exxon's earliest successful applications involved investment planning under economies of scale. Operations scheduling applications are characterized by mutually exclusive variables: these have been solved satisfactorily by heuristic methods and by branch-and-bound methods running under streamlined computational procedures. Nonconvex constraints are found in engineering design problems: these require artful formulation and specialized computational search procedures. Research is still needed to endow discrete mathematical programming with interactive computation capabilities, with enhanced analytical and interpretive options and with extensions into the domain of mathematical programming under uncertainty.

1. Beginnings. To introduce this topic, I can do no better than to paraphrase a 1958 RAND Corporation paper [1] (see also [2]) by George Dantzig. Figure 1 summarizes part of his paper which enumerates a variety of uncomputable problems which would become computable if mixed-integer programming proves to be as successful as linear programming.

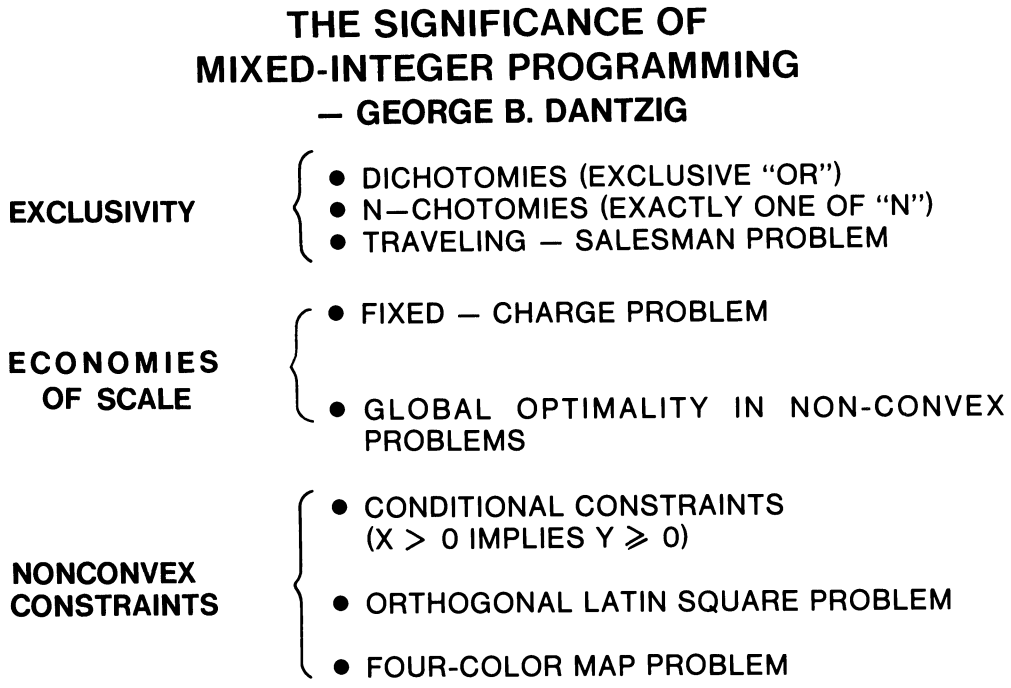


FIG. 1

* Received by the editors July 27, 1981.
† Communications and Computer Sciences Department, Exxon Corporation, Florham Park, New Jersey 07932.

Although Dantzig's paper gives the appearance of optimism, I suspect that he wrote it with tongue in cheek. I think he was warning that it would be most remarkable if a new computational technique would serve to sweep away, at a single stroke, so many problems which have baffled the best mathematical minds for decades or even centuries. Experience over the 23 years since his paper's publication has shown Dantzig's caution to have been well founded.

The subdivision of Dantzig's list into three rough categories is my own; it is intended mainly to provide some perspective on the nature of the underlying mathematical structures of the various applications he discusses. "Exclusivity" arises, for example, in ship scheduling: one cannot send parts of a ship to two or more separate ports without adversely affecting its buoyancy.

Many of our applications, particularly under economies of scale in investment planning, display nonconvexity only in the *objective function*. The essential nature of the computation is the search for a global optimum among the various local peaks. In such problems, there is generally no difficulty in moving through a sequence of feasible solutions while travelling from one local optimum to another.

Nonconvexity in the *constraint set* arises either out of restrictions coming from the physical sciences or out of regulations devised by the cunning of bureaucrats. I will provide examples of each. Problems of this sort tend to be among the most computationally difficult.

1958 was a seminal year in several respects for discrete variable mathematical programming at Exxon. In addition to the Dantzig paper I have mentioned, the Gomory cutting plane papers [3], [4], [5], [6], began to appear, and in our own offices, we began to give serious thought to one of our first plant-scheduling applications.

Figure 2 suggests some of the options available to a scheduler in a plant of this kind. (This particular plant is a product of my own imagination.) The scheduler's primary objective is to maintain an adequate inventory of the various products in the tanks on the right, so that all deliveries into final demand can be satisfied promptly. To do this he or she can operate the various processes in different ways and on different feedstocks. The fundamental nonconvexity of the problem arises out of the fact that one can operate a given process in only one fashion at a time. This is completely analogous to my previous statement that one cannot send fractions of a ship separately to different ports at the same time. An additional element of complexity is introduced by the fact that the processes are not always decoupled by intermediate tankage; in the diagram, process D can operate only if process C or process E (or both) is simultaneously scheduled to receive and handle its product stream.

Operation I in Fig. 2 shows one possible way in which this plant might operate during a given interval of time. Notice first that deliveries into final demand always take place out of the product tankage on the right. This diagram suggests that, considering the inventory levels shown, operation I may be terminated by any one of three possible events:

- Tank 9 may run dry;
- Tank 6 may run full;
- Tank 3 may run dry.

When an event such as these occurs, the operator must necessarily switch to some other operation. Another point to note is that because of the absence of an intermediate tank in this line process D and process C are directly linked in throughput rate. We suppose that these rates are, in fact, compatible, and that each of the processes shown

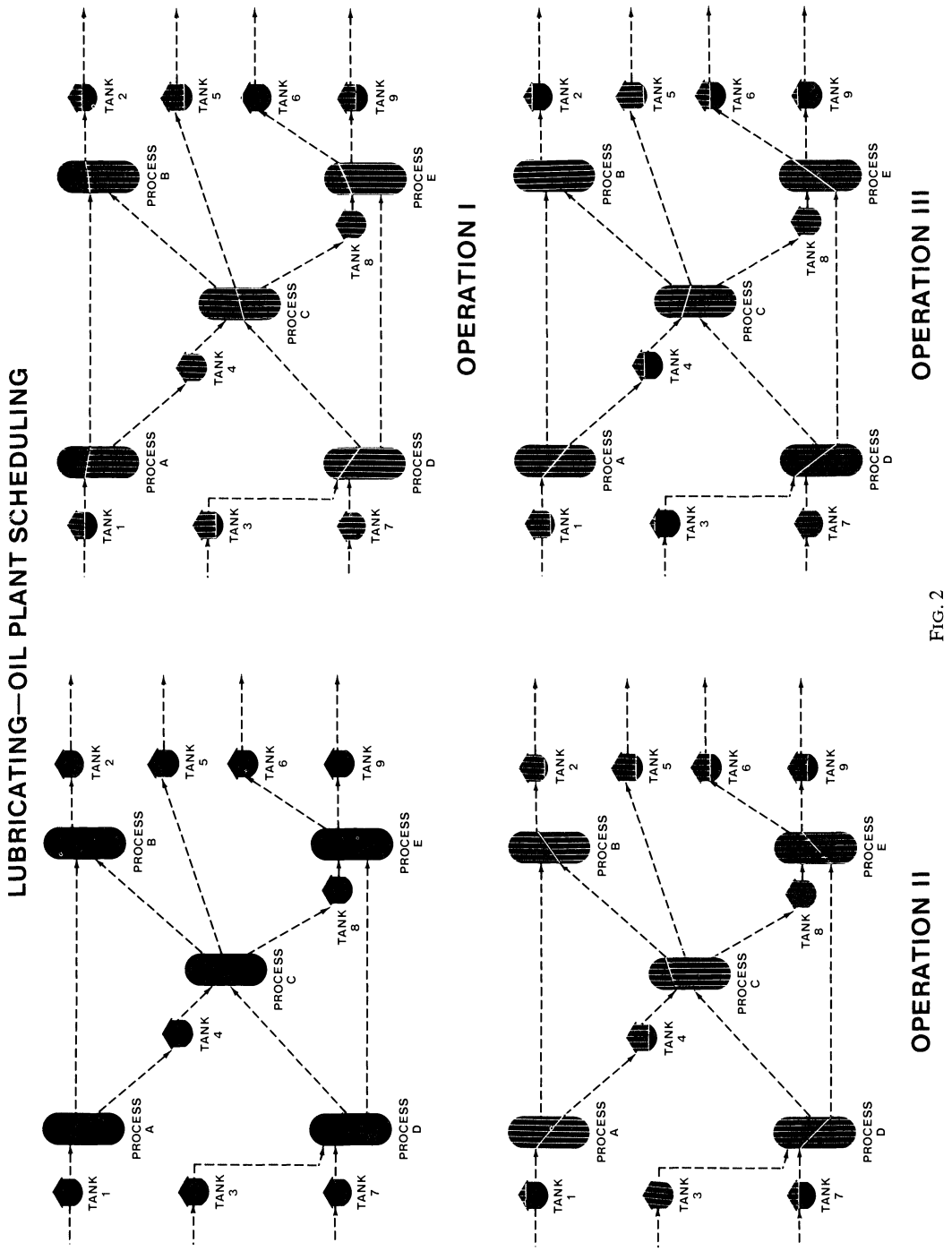

LUBRICATING—OIL PLANT SCHEDULING
OPERATION I
OPERATION II
OPERATION III
TANK 1
TANK 2
TANK 3
TANK 4
TANK 5
TANK 6
TANK 7
TANK 8
TANK 9
PROCESS A
PROCESS B
PROCESS C
PROCESS D
PROCESS E

FIG. 2

does, in fact, produce product of appropriate quality to be delivered into the particular grade of final product tankage. Such capacity constraint and product quality restrictions are a part of the "continuous" portion of a mixed-integer formulation; they do not generally exhibit any nonconvexity properties of their own.

Operation II is an alternative operation in which there is some degree of interdependence of process throughput rates. This also suggests that there is more than one way to meet the quality specifications for some of the final products.

Operations I and II suggest, by the status of the final product inventory levels, that the scheduler is usually under pressure to get the maximum throughput out of the plant in order to keep up with the rate of deliveries into final demand. Since a process unit generally goes through an unproductive period during changeover from one operation to another, capacity in effect is lost during changeover. Thus a so-called "switching cost" is a part of the objective function. Other cost considerations may arise from the possibility of producing the same product from different raw materials or by means of different processes. These considerations result in what may be termed the "convex" aspect of the scheduler's objective function.

In Operation III, a measure of "decoupling" is achieved by temporarily idling process B. Notice that full "decoupling" could be achieved (in the sense that the throughput rate of each operating process unit could be set independently of all others) by idling process D and feeding process E out of tank 8. Other things being equal, the scheduler will generally prefer such a "decoupled" operation, partly because it is easier to schedule and partly because the units are easier to control when their feed rates may be set independently. For the same reason, the scheduler will generally press for more intermediate tankage than is shown in my example. The analysis of the economic justification of such tankage is an interesting exercise in that it weighs the "hard" cost of tankage against the "soft" incentives of easier scheduling and improved process control, with a somewhat nebulous potential for increased throughput. (Recall that we attained "decoupling" at the cost of idling two of the five process units, which certainly implies a sacrifice in potential plant throughput.)

When we worked on this problem in 1958 (see Fig. 3), our kit of ready-made tools consisted of only economic-order-quantity analysis and linear programming. Nevertheless, we were able to establish most of the solution properties which are standard today:

- A trial schedule consists of a sequence of "operations," together with a duration time for each.
- The objective function is obviously a nonconvex function of the "operations" *sequence*, and less obviously a nonconvex function of the *duration times*.

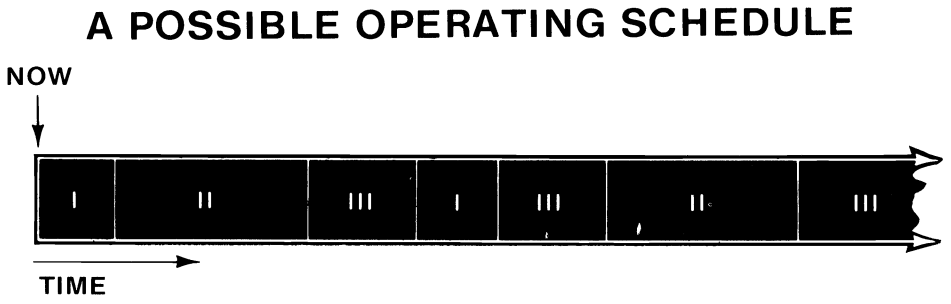


FIG. 3

- Under conditions of sufficient regularity in demand patterns, and of some degree of commensurability of the data, some sort of "limit cycle" schedule can be found which can be used over and over again into the indefinite future.
- When the demand is less regular or stochastic, it would appear that the "limit cycle" could still play a role as a target toward which ad hoc scheduling should constantly aim to move.

None of these insights were sufficiently concrete for use in the actual plants in 1958.

In 1960, we chanced upon an application which, although it never came to practical fruition, served to call forth essentially the same three solution techniques which we see in practical use today. See Fig. 4. At that time (and I am pleased to be able to say that this is no longer the case) the Canadian province of Alberta imposed a most ingeniously devised set of rules for bidding on potential crude-producing leases. The purpose was to reserve for public ownership and/or for possible later sale into the private sector a significant fraction of the most promising lands for crude production. However, this reservation had to be done before actual drilling had shown where the most promising lands were. On a 6-mile by 12-mile tract, broken up into $\frac{1}{2}$-mile squares as illustrated in Fig. 4, the following bidding rules were imposed:

1. The total bid must include no more than 50% of the squares in the tract.
2. Each block of contiguous squares must be square or rectangular in shape. It must not exceed 36 squares in area, and its longer dimension must be no greater than twice its shorter dimension. Block boundaries must fall on the square boundaries in the diagram.
3. Each leased block must be separated from other leased blocks by at least two rows of squares, except in cases where two blocks touch only at their corners (checkerboard).

The bidder's interest, obviously, is to pick a pattern of blocks which has the highest possible potential for the discovery of crude reserves, as determined by whatever geological data may be at hand. The dark contours on this chart represent the current state of the bidder's estimates as to where oil is likely to be found.

Figure 4 also shows a plausible bid pattern (which I generated by hand analysis). Two points are worth noting about this:

- If the contour pattern is sufficiently smooth and slow moving in its major patterns, then the province will have reserved for itself broad corridors into the interior of the promising areas.
- The selection of the most worthwhile blocks is obvious, but the "fine tuning" of the pattern to attain optimality can be tedious and unrewarding.

I mentioned earlier that this application called forth the three main classes of solution procedures: the man-machine interactive, the algorithmic and the heuristic procedures.

On the man-machine interactive level, it is obvious, after a little playing around with this problem, that the two main factors limiting the hand solution process are the tedium of redrawing each successive trial bid pattern on paper plus the inconvenience of summing up the expectations for 144 squares at each trial. Today one could easily provide these services via a personal computer. But remember that the time was 1960; this was a clearcut case of a research idea born at the wrong time. If we had it to do over again in the present day of personal computing and sophisticated

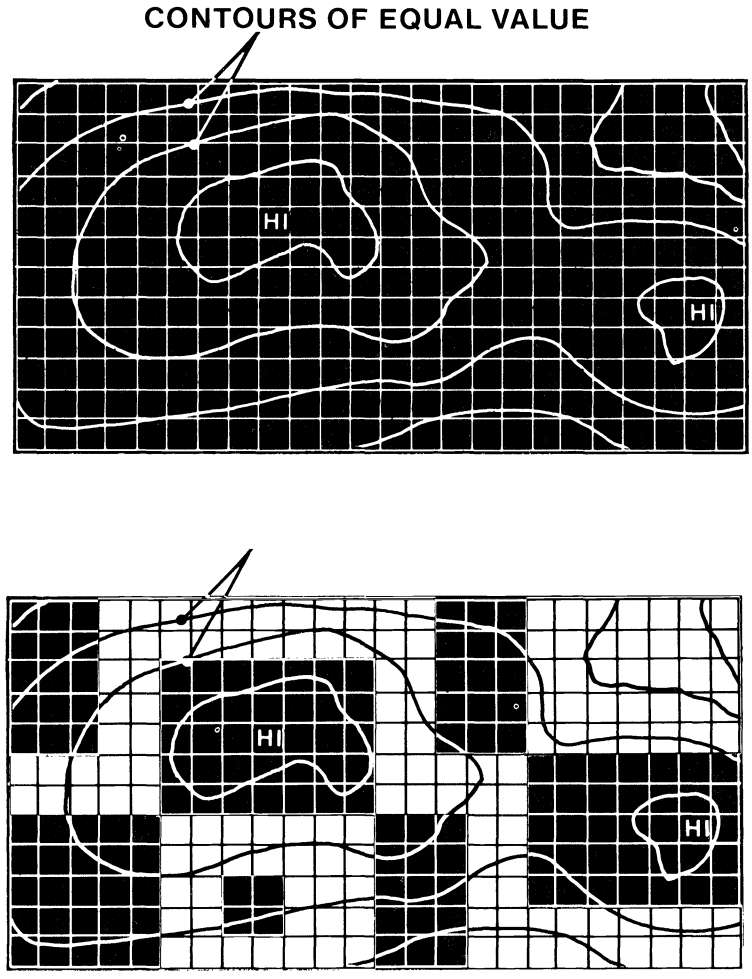


FIG. 4

graphics terminals, I think that this approach might well have carried the day. In 1960 the required development time just didn't fit the "acceptability" window.

It may surprise the reader to learn that we could code a Gomory cutting plane procedure faster than we could set up an interactive graphics terminal, but the fact was that we were already into the cutting planes research. Here again the result was that computers had not developed sufficiently to make this approach feasible. We found that we had to reduce the problem size drastically by making the spacing of the square boundaries only half as dense, in order to get a solution from the computer. Such a solution is of "academic interest" only (which is the same as "no interest at all" to the operating people).

The heuristic approach *evolved* (as heuristic approaches are wont to do) from the observation that *evaluation* of a proposed bid pattern was an order of magnitude

faster than the process of *algorithmic generation* of (supposedly) improved patterns. Thus the proposal: "Why not generate a large population of proposed patterns quickly by a random process, rather than a small population slowly by a more analytical process?" Experience with this approach showed that it was rather poor at building up a good pattern "from scratch," but that it could quite consistently improve on the best hand solutions submitted to it as a starting point. Proof of optimality is, of course, out of reach via this approach.

A final comment on the class of applications represented by this example (that is, the problems involving nonconvex constraints) is that a good deal of craftsmanship is required to arrive at a formulation which is maintainable, understandable and computable. In this class of problems, we do not foresee an early passage to the "black box" type of usage which constitutes a large part of linear programming practice today.

Up to this point, I have been discussing applications which did not "fly" in the sense of being put to use by the operating people. The next application proved to be a real moneymaker—in fact, its lineal descendents are still doing good work.

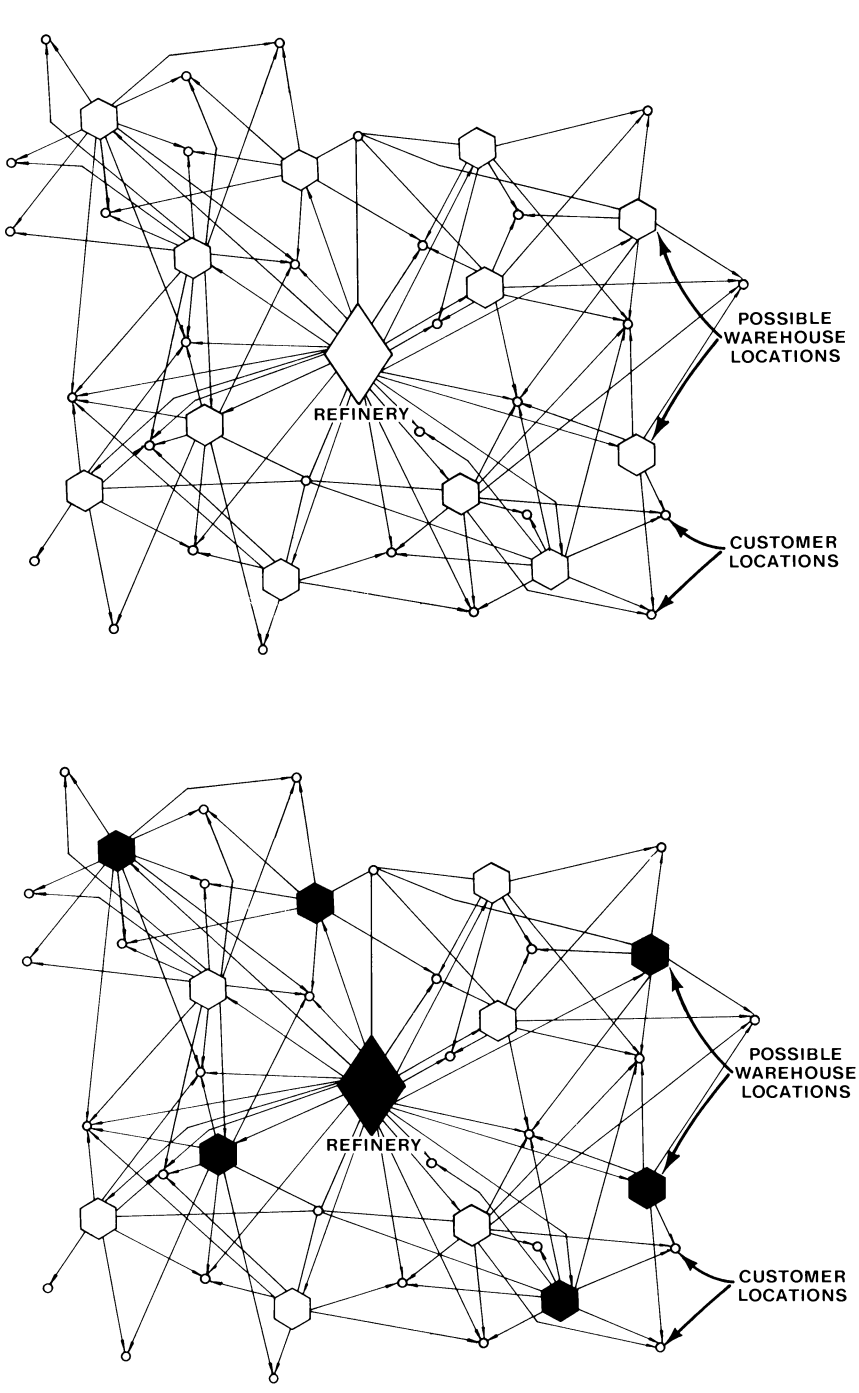
To achieve economical distribution and good customer service in such markets as heating oil and motor gasoline, it is necessary to ship the products in large carriers to local distribution points (which we will call "warehouses") and to move them from there to the final customer in smaller carriers such as short-haul multi-drop tank trucks. The question is, "how many such warehouses should there be, and where should they be located?" The answer involves an economic balance between the distance-dependent costs of transportation and the (mostly fixed) costs of establishing warehouses at various possible sites. Figure 5, although it greatly understates the number of sites and routes, nevertheless conveys some impression of the complexity of the problem.

We used branch-and-bound procedures as proposed by Land and Doig [7] as a conceptual basis for this computation. The problem lends itself to this approach because the $0-1$ discrete variables are each associated with the existence or non-existence of a particular warehouse. Once a particular "branch" is chosen, i.e., a particular subset of warehouses is selected to be open or closed, what remains is a classical network problem, which can be solved very rapidly. (In any branch-and-bound procedure, a great many variants of the continuous subproblem must be solved thus, so it is important that these subsolutions be obtained rapidly and cheaply.)

Figure 5 also serves to illustrate one peculiarity of branch-and-bound procedures: under some branch selections, the continuous subproblem may have no feasible solution. In particular, there is no way to supply the customer in the lower left-hand corner once his one and only warehouse is shut down.

2. Attempts at a general purpose code. I think that these three examples, one from the scheduling area, one from the nonconvex constraints area and one from the economies-of-scale area, will suffice to give a general flavor of the discrete variable applications we generally see. We thought in the early 1960's that we had done enough special purpose work in the field, and were thus qualified to produce a general purpose mixed-integer code. The result was MISTIC, a code for the IBM 360, which appeared in 1964. See Fig. 6. In the ensuing five years we attempted various applications:

1965: Investment planning under fixed charges;
1965: Compressor-driver selection;
1968: Pipeline optimization;
1969: Investment planning under continuous economies of scale.



ROUTING OPTIONS UNDER ONE PLAUSIBLE WAREHOUSE SELECTION

FIG. 5

I think it is fair to regard all of these as variants of the economies-of-scale problem. Certainly process scheduling is conspicuous by its absence. It is significant, I think, that all these attempts were from the class of problems which is probably easiest to compute, which has the largest economic incentive and which has the least pressing deadlines for implementation of a solution. The fact was that MISTIC was implemented within the framework of a general purpose LP code. Within this framework, each new "branch" operation meant paging the whole linear programming apparatus out of memory, making some minor changes in the problem data, and then paging the whole apparatus back in again. The logistics of data movements within the computer system tended to bog down the whole process.

MISTIC: A 1964 (IBM 360) CODE

1965 : INVESTMENT PLANNING UNDER FIXED CHARGES

1965 : COMPRESSOR-DRIVER SELECTION

1968 : PIPELINE OPTIMIZATION

1969 : INVESTMENT PLANNING UNDER CONTINUOUS ECONOMIES OF SCALE

FIG. 6

3. 1970–1980: Code expediting and application insights. During the 1970's, we gradually built up a much more confident posture in our mixed-integer capabilities. I think this is attributable to three factors. First, at the data logistics level within the code, we have, with KETRON'S help, managed to develop procedures for making the required minor changes in problem data without paging the whole system in and out. This has resulted in great savings in computing time. See Fig. 7.

The second factor is the discovery of means by which the user's insights into the problem may be used to expedite the tree search in branch-and-bound. Essentially, this is done by allowing the user to convey information about the model's structure to the branching algorithm via the naming conventions used for the variables.

The third factor could be regarded as a "sour grapes" philosophy: we often decide to be content with a "good" solution which has not been proven computationally to be optimal. In both of these latter factors, one sees something of the heuristic philosophy which was so successful in early applications. I personally do not believe that this means a permanent abandonment of the optimization goal; optimal solutions offer side benefits of sensitivity analysis and case comparability which cannot be ignored. But I think we are saying that we would rather settle for a "satisfying" solution than lose the interest of our users.

With the streamlined capabilities of BLOODHOUND, we are making the rounds of the traditional applications once more (see Fig. 8) and, I think, getting a little further with each of them.

The scheduling type applications are still approached in two ways: heuristic methods such as materials requirements planning, and algorithmic methods such as BLOODHOUND [6]. (As I suggested earlier, the differences between these methods are not as extreme as one might suppose; research in MRP is moving toward enhanced

1970: BLOODHOUND = EXPEDITED BRANCH & BOUND

- FACILE REDEFINITION OF CONTINUOUS NON-CONVEX PROBLEMS
- FLEXIBLE SPECIFICATION OF TREE-SEARCH STRATEGIES
- OFTEN TERMINATED SHORT OF PROVEN OPTIMALITY

FIG. 7

algorithms, and research in mixed-integer methods is moving toward increased heuristic use of user insights. Both are talking about moving into interactive computation in response to user needs.) At present, we recommend MRP where the number of potential processes is large and the degree of interaction among processes is small. Conversely, we recommend MIP when the number of processes is small and the degree of interaction among processes is large.

Economies-of-scale problems offer no particular difficulty except when the time dimension is important. I will return to this point in a discussion of research frontiers.

Dealing with constraint nonconvexities is still a craft in formulation and an art in computation. In general, one must settle for a guarantee only of *local* optimality, but any degree of confidence about global optimality must be derived from a priori knowledge about the application itself.

APPLICATIONS REVISITED

- SCHEDULING-TYPE
  - + MATERIALS SUPPLY INVENTORIES
  - + SHIP SCHEDULING
  - + PROCESS SCHEDULING
- ECONOMIES OF SCALE
  - + CAPITAL BUDGETING
  - + FACILITIES PLANNING
  - + EQUIPMENT CONFIGURATION
- CONSTRAINT NONCONVEXITIES
  - + POOLING OF HYDROCARBON STOCKS
  - + PROCESS OPTIMIZATION
  - + DEVELOPMENT OF CRUDE RESERVOIRS

FIG. 8

4. Gaps in our knowledge. Being a coordinator and manager of our research program in computer-based management tools, I find it impossible to be contented with the present state of our knowledge in this field. Figure 9 shows some of the areas in which, I believe, our tools need to be improved.

GAPS IN OUR KNOWLEDGE

- INTERACTIVE CALCULATIONS
- DISCRETE VARIABLES IN CONTINUOUS TIME
- PERTURBATIONS AND SENSITIVITY — ANALYSIS
- OPTIMUM CHOICE OF DISCRETE VARIABLES UNDER CONTINUOUS-TIME STOCHASTIC PROCESSES

FIG. 9

I have already mentioned our initiatives in the direction of man-machine interaction. Let me make it clear that I do not expect, in general, that such techniques will lead to cheaper ways of computing optimal solutions. Rather, I hope to see the emergence of better insights and understanding on the part of the user as to the fundamental nature of the application. Too often in the past decade we have been so algorithm-minded that we have encouraged the user to deal with the computer on a "black box" basis. This leads to a degradation of user competence and to a rapid evaporation of management confidence in the whole operation.

I mentioned the problem of embedding discrete variable formulations in continuous time. I know enough about convex formulations in continuous time to realize how much of the true structure of the optimal solution is obscured by the discretization of the time dimension. It boggles my mind to think how much distortion we cause by introducing a new integer variable for the same decision option in each successive time period! Furthermore, such a practice in an n-period model increases the number of branches in the tree search by a factor of about 2^n.

I contend that the greatest power of linear programming lies in its analytical capabilities—for example, the interpretation of the dual solution, the interpretation of the inverse matrix and the parametric capabilities. As a close relative of LP, MIP could have similar capabilities. But these potentialities are rarely mentioned, let alone exploited.

The last point of this paper is also my foggiest. I assume we are all familiar with Dantzig's model for multistage planning under uncertainty. But have we considered how the structure of the solution will be altered if some of the variables in each period are restricted to integer values? We all know a few folk theorems pertaining to discrete variable decision making problems under uncertainty:

- The planning horizon under uncertainty is closer than under the corresponding certainty case.
- The effective discount factor under uncertainty is smaller (i.e., implying a higher effective interest rate) than under the corresponding certainty case.
- The identification of the discreteness of integer variables becomes less and less necessary in remoter time periods in the uncertainty case, whereas it does not fade out at all in the certainty case.

May I hope that someone in the reading audience will clothe these ideas with rigor, preferably within the time-span of my professional career?

REFERENCES

[1] GEORGE B. DANTZIG, *On the significance of solving linear programming problems with some integer variables*, Paper P-1486, RAND Corporation, 1958.

[2] ———, *On the significance of solving linear programming problems with some integer variables*, Econometrica, 28 (1960), pp. 30–44.

[3] R. E. GOMORY, *Essentials of an algorithm for integer solutions to linear programs*, Bull. Amer. Math. Soc., 64 (1958), pp. 275–278.

[4] ———, *An algorithm for integer solutions to linear programs*, Tech. Rep. 1, Princeton-IBM Mathematics Research Project, November, 1958.

[5] ———, *An algorithm for the mixed integer problem*, Paper P-1885, RAND Corporation.

[6] ———, *Extension of an algorithm for integer-solutions to linear programs*, Abstract 553-190, AMS Notices, 6, 1 (1959), p. 52.

[7] A. H. LAND AND A. G. DOIG, *An automatic method of solving discrete programming problems*, Econometrica, 28 (1960), pp. 497–520.

[8] THOMAS E. BAKER, *A branch and bound algorithm for interacting process scheduling*, Math. Programming, Special Studies Issue 15 (1981), pp. 43–57.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0007 $01.00/0

SORTING AND MERGING IN ROUNDS*

R. HÄGGKVIST† AND P. HELL‡

Abstract. The need for sorting algorithms which operate in a fixed number of rounds (rather than have each new comparison depend on the outcomes of all previous comparisons) arises in structural modeling. Since all comparisons within a round are evaluated simultaneously, such algorithms have an obvious connection to parallel processing.

In an earlier paper (SIAM J. Comput., 10 (1981), pp. 465–472) we used a counting argument to prove the existence of subquadratic sorting algorithms for two rounds. Here we develop optimal algorithms for merging in rounds, and apply them to actually construct good sorting algorithms for k rounds, $k \geqq 3$. For example, in $k = 66$ rounds, our algorithm will sort any n-element linearly ordered set with $O(n^{1.10})$ comparisons.

1. Motivation. We shall consider two sources of our problem, one in system modeling and one in theoretical computer science.

Interpretive structural modeling (ISM) is a technique, developed by J. Warfield at Battelle Memorial Institute [14]–[17], to provide complex systems with structure. Typically, this takes the form of introducing a binary ("contextual") relation on the elements of the system and displaying the relation in some easily understandable form. The first important step of ISM (after the elements of the system have been identified and the contextual relation decided upon) consists of forming a "matrix model" of the system, i.e., the characteristic matrix M of the relation R. (The i, jth entry of M is 1 if the ith element is related in R to the jth element, 0 otherwise.) In this step we pass from a "mental model" of R (in the mind of a subject or as consensus of a group of subjects) to the matrix model M. In consequent steps of ISM, not of interest in this paper, one takes the matrix model M to a "hierarchical digraph" D and perhaps to a minimal digraph D' with the same transitive closure as D. Displaying D or D' as a hierarchy is then useful in the analysis of the complex system. We are interested in the formation of the matrix model. When the relation R is transitive, such as preference is usually assumed to be (see [17, p. 295] for a discussion of possible objections to assuming transitivity in practice), the construction of the matrix M is facilitated by transitive inference. Warfield [14]–[16] presents an algorithm to construct M which prompts the subject (or group of subjects) possessing the mental model of R with queries of the type "Is x related in R to y?" (for instance, "Do you prefer x to y?" or "Does x impact on y?") Depending on the answer, the algorithm fills in at least one entry in M (and perhaps more if transitivity can be employed) and decides the next query to be posed. At the end of this process the matrix M will have been formed. Clearly, if R is a linear order (or a weak order [9]), the same objective can be achieved by any of the sorting algorithms based on binary comparisons [7], [8]. (In fact, sorting the elements will yield directly the hierarchical digraph, [17].)

The techniques of ISM have been widely applied [3], [6], [12], [13]. A possible drawback of the existing methods of ISM, as far as the construction of M is concerned, lies in the necessity of submitting the subject (or group of subjects) to a session at a computer terminal. In some instances the formation of M must be done by correspondence. (For example, that was the case in a recent study of preference among environmental alternatives [10] where the subjects were representatives of various

* Received by the editors June 11, 1981.

† Institute Mittag-Leffler, Djursholm, Sweden.

‡ Department of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6.

U.S. organizations concerned with the environment.) In such a situation a natural way to form M is to pose a number of queries (binary comparisons) simultaneously, record the answers together with any entries implied by transitivity and evaluate the result to decide on the next set of queries; this process might continue for a number of rounds, not to exceed a given integer k. We shall describe a method by which this can be achieved and which is guaranteed to make a reasonably small number of comparisons.

In what follows we assume that the relation R is a linear order. This corresponds to strict preference [9] (i.e., the subject is not allowed to be indifferent between two alternatives), and the digraph D depicting R is a transitive tournament. It is not difficult to see that our results hold when R is a weak order [9]; this is the case of weak preference, i.e., when indifference is allowed. In this case D is a digraph whose condensation is a transitive tournament and whose strong components are complete digraphs (corresponding to the sets of alternatives amongst which the subject is indifferent).

The fact that our technique depends on evaluating several queries simultaneously (in parallel) results in an obvious alternative interpretation as an algorithm for several parallel processors. Parallel processing in comparison problems has enjoyed considerable interest [7], [11]; our algorithms are interesting in that they guarantee that only a constant amount of time will be spent making comparisons. It should be noted, however, that we ignore not only the time for operations other than binary comparisons, but also questions of communication among the processors and between the processors and the outside environment. Nevertheless, our results are of interest when analyzing the tradeoff between the time needed for parallel comparisons (i.e., the number of parallel steps) and the available number of processors.

2. Parallel comparison trees. Because of the differences in the two sources of our problem, we encounter a difficulty with terminology. We resolve it by resorting, for the most part, to the language of computer science. Thus, the elements are usually referred to as keys, queries as binary comparisons and the construction of the matrix M as sorting. Furthermore, we use a generalization of comparison trees [7] for the description of our algorithms; a *parallel comparison tree* allows a number of (not necessarily disjoint) binary comparisons at each node (cf. Figs. 1 and 2). This is analogous to the way binary comparison trees are used to describe and analyze ordinary sorting algorithms [7]: the computation begins at the root with evaluating the comparisons indicated; depending on the outcomes a corresponding branch is taken leading to the next set of comparisons to be made and so on until a leaf is reached yielding a sorted sequence of the keys (or equivalently a matrix of the relation $<$). The height of the tree is the time spent making comparisons (in the worst case), i.e., the greatest

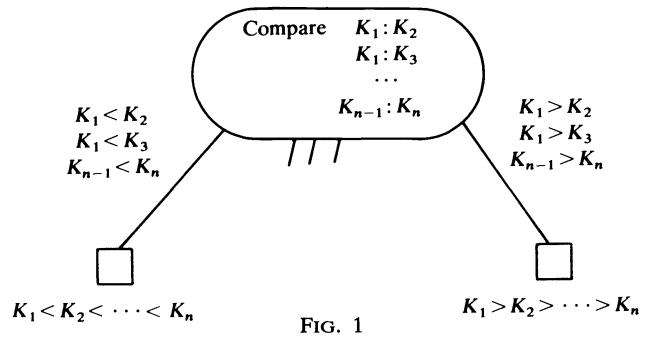


FIG. 1

number of parallel steps. Note that parallel comparison trees need not be binary trees. Figure 1 illustrates the obvious fact that with $\binom{n}{2}$ processors all $\binom{n}{2}$ comparisons can be performed in one time interval.

Let for a parallel comparison tree T the maximum total number of comparisons in any root-to-leaf path in T be cp (T) and the maximum number of comparisons in any node of T be cn (T). We define SORT $(k, n) = \min$ cp (T), where the minimum is taken over all parallel comparison trees T of height k which sort n keys; similarly, SORTP $(k, n) = \min$ cn (T) over the same set of trees T. (Figure 1 implies that

$$\text{SORTP}(1, n) \leqq \text{SORT}(1, n) \leqq \binom{n}{2}.)$$

Evidently, SORT (k, n) is the minimum number of queries needed to guarantee that for any linear order R on any set of n elements the matrix M can be formed in k rounds. Similarly, SORTP (k, n) is the minimum number of processors needed to assure that any set on n linearly ordered keys can be sorted by binary comparisons so that all comparisons are arranged to take place in a constant time of k intervals.

In what follows we shall state all results in terms of SORT (k, n). Unless stated otherwise, all evaluations, lower and upper bounds, apply to SORT (k, n) as well. This is due, for the most part, to the fact that k is fixed and that our algorithms have about the same worst case number of comparisons at each level of the tree.

As we stated above it is obvious that SORT $(1, n) \leqq \binom{n}{2}$ (cf. Fig. 1): it is almost as obvious that SORT $(1, n) = \binom{n}{2}$, since if not all pairs of keys are compared at the root node of T, then there exists a child of the root in which the order is not completely determined and T must have height strictly greater than 1.

A parallel comparison tree of height 2 which sorts n keys is completely described by its set of comparisons at the root node. Indeed, it is again easy to see that in any child of the root all comparisons not made at the root or implied from the answers by transitivity will have to be made. Thus, we can identify such trees with undirected graphs on the set of keys (cf. Fig. 2). Conversely, given such a graph G, the comparisons indicated by the edges of G are performed first (at the root node), the result being an acyclic orientation of G.

After the transitive closure of the orientation has been taken, certain pairs of keys may remain incomparable and they are compared at the corresponding node in the second level of the tree. For instance, in Fig. 2 each node other than the root contains at most 4 comparisons, which illustrates the fact that SORT $(2, 5) \leqq 9$ and SORTP $(2, 5) \leqq 5$. (In fact, it is simple to verify that SORT $(2, 5) = 9$ and that

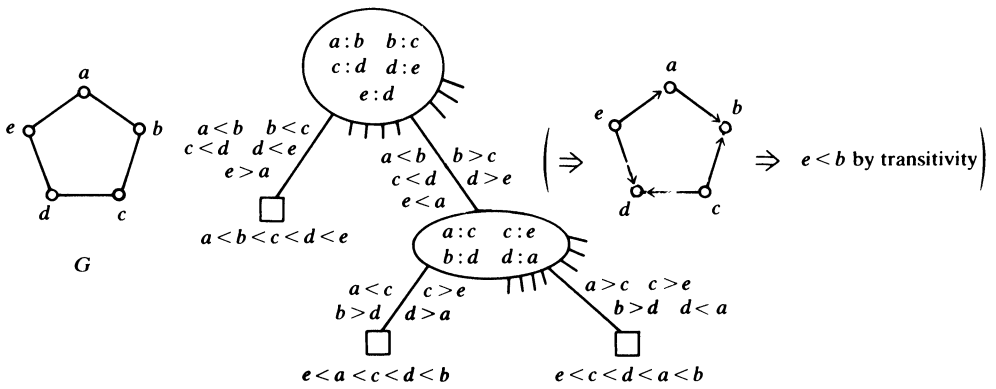


FIG. 2. *The parallel comparison tree of height* 2 *corresponding to the given graph G.*

SORT $(2, n) = \binom{n}{2}$ for $n < 5$; similarly, SORTP $(2, n) = \lceil \frac{1}{2}\binom{n}{2} \rceil$ for $n < 6$ so that SORTP $(2, 5) = 5$.) Thus, we can view SORT $(2, n)$ as the minimum over all graphs G with n vertices of the sum (respectively, maximum for SORTP $(2, n)$) of the following two quantities A and B: A is the number of edges of G; B is the maximum over all acyclic orientation of G of the number of edges missing from the transitive closure of the orientation. In other words, one seeks a graph which is sparse (has few edges) but such that each acyclic orientation has transitive closure which is dense (has many edges). We have proved in [4] that there exist graphs G with n vertices and $O(n^{5/3} \log n)$ edges which are guaranteed to have only $O(n^{5/3})$ edges missing in the transitive closure of any acyclic orientation. (We have not actually constructed such graphs; in fact, it seems hard to find graphs with $o(n^2)$ edges guaranteed to have only $o(n^2)$ edges missing in the transitive closure of any acyclic orientation [2].) Thus, SORT $(2, n) \leqq O(n^{5/3} \log n)$. We have also proved[1] that SORT $(2, n) = \Omega(n^{3/2})$ and extended our results to

$$\Omega(n^{1+(1/k)}) = \text{SORT}(k, n) = O(n^{\alpha_k} \log n)$$

for every fixed k. Here $\alpha_k = (3.2^{k-1} - 1)(2^k - 1)$ so that $\lim \alpha_k = \frac{3}{2}$ [4]. Inasmuch as our results depended recursively on the graphs G above, which have not been constructed, the upper bound must be viewed only as an existence result. (There exists an algorithm—a parallel comparison tree—of the given complexity, but we have not constructed it.) Although we still do not have a construction for the graphs G, we shall use parallel comparison trees for merging and apply the results to SORT (k, n) for $k \geqq 3$. In fact, we shall find a sequence s_k with $\lim s_k = 1$ and improve the upper bounds to

$$\text{SORT}(k, n) = O(n^{s_k})$$

for each fixed k. (Thus, the exponents of n in both the upper and lower bounds of SORT (k, n) have the same limit.)

A number of results concerning order statistics by parallel comparison trees may be found in [5].

3. Merging. We shall assume as given two subsets of size n of a linearly ordered set and use parallel comparison trees T of fixed height k to merge the sets together (i.e., to sort the union of the two sets). Let MERGE (k, n) denote the minimum of cp (T) over such trees T and MERGEP (k, n) the minimum of cn (T) over the same set of trees T. As before, all results cited for MERGE (k, n) apply to MERGEP (k, n) as well.

For trees of height 1, it is easy to see that at the root node each key in one set must be compared to each key in the other; thus, MERGE $(1, n) = n^2$.

THEOREM 1. MERGE $(2, n) = O(n^{4/3})$.

We give a constructive proof of the theorem, i.e., describe parallel comparison trees T of height 2 which merge two linearly ordered sets of size n with $O(n^{4/3})$ comparisons. This will again be best done by describing the (bipartite) graph consisting of all comparisons performed in the first round.

Let $A = \{a_1, a_2, \cdots, a_n\}$ and $B = \{b_1, b_2, \cdots, b_n\}$ be the two linearly ordered sets, written in increasing order (i.e., $i < j$ implies $a_i < a_j$ and $b_i < b_j$). Let G be the undirected graph with the vertex-set $A \cup B$ and the edges $a_i b_j$ for all i and j divisible by $f = \lfloor n^{1/3} \rfloor$.

[1] For functions f, g with nonnegative values $f(n) = \Omega(g(n))$ means $f(n) \geqq c.g(n)$ for some c and all sufficiently large n.

An orientation O of G is called *admissible* if the digraph O' consisting of O together with the arcs from a_i to a_j and b_i to b_j for all $i<j$ is acyclic. The *extended transitive closure* of O is defined to be the transitive closure of O'.

Let $g=\lceil n/f\rceil$. The set A is partitioned into $A_1=\{a_1,\cdots,a_f\}, A_2=\{a_{f+1},\cdots,a_{2f}\},\cdots,A_g=\{a_{(g-1)f+1},\cdots,a_n\}$; similarly for B. For each admissible orientation O of G, we define an undirected graph G^* with the vertex-set $\{A_1,A_2,\cdots,A_g\}\cup\{B_1,B_2,\cdots,B_g\}$ and the edges A_iB_j for all i and j such that in the extended transitive closure of O there exist $a\in A_i$ and $b\in B_j$ not adjacent in either direction.

LEMMA 1. *For any admissible orientation of G, the graph G^* is planar.*

Proof. In fact, we prove that if $i<i'$ and $j'<j$, then it is not possible for both A_iB_j and $A_{i'}B_{j'}$ to be edges of G^*. Consider the edge $a_{if}b_{(j-1)f}$ of G. If O contains the arc from a_{if} to $b_{(j-1)f}$, then for any $a\in A_i$ and $b\in B_j$, O' contains the directed path $a, a_{if}, b_{(j-1)f}, b$, and hence, in the extended transitive closure of O, each $a\in A_i$ is adjacent to each $b\in B_j$; therefore, A_iB_j is not an edge of G^*. On the other hand, if O contains the arc from $b_{(j-1)f}$ to a_{if}, then for any $a\in A_{i'}, b\in B_{j'}$, O' contains the directed path $b, b_{(j-1)f}, a_{if}, a$, and hence, in the extended transitive closure of O, each $a\in A_{i'}$ is adjacent from each $b\in B_{j'}$; thus, $A_{i'}B_{j'}$ is not an edge of G^*. (Note that the lemma is valid for any choice of f.)

The corresponding algorithm (parallel decision tree) for merging A and B in two rounds can now be described as follows: In the first round make all comparisons named by the edges of G. This results in $\lfloor n/f\rfloor^2=O(n^{4/3})$ comparisons. After these comparisons have been evaluated, we obtain an admissible orientation O of G. All the comparisons whose results are not in the extended transitive closure of O will be performed in the second round. If the comparison between some $a\in A_i$ and some $b\in B_j$ needs to be made in the second round, then A_iB_j is an edge of G^*. Moreover, each edge of G^* represents at most $\lfloor n^{1/3}\rfloor\cdot\lfloor n^{1/3}\rfloor=O(n^{2/3})$ comparisons in the second round. Since G^* is planar, it has only $O(g)=O(n^{2/3})$ edges. Consequently, there are only $O(n^{4/3})$ comparisons made in the second round.

We can now describe parallel comparison trees of height k, which merge two sets of size n.

COROLLARY 1. *For each fixed k,* MERGE $(k,n)=O(n^{\beta_k})$, *where* $\beta_k=2^k/(2^k-1)$.

Proof. We proceed by induction on k starting with $k=2$ and Theorem 1. (MERGE $(1,n)=n^2$ is obvious.) Let $\gamma_k=(2^{k-1}-1)/(2^k-1)$ and note that $\gamma_k\cdot\beta_{k-1}=1-\gamma_k$ and $2-2\gamma_k=\beta_k$. Construct a graph G as in the theorem but with $f=\lfloor n^{\gamma_k}\rfloor$ (and, hence, $g=O(n^{1-\gamma_k})$) and proceed as before. In the first round compare all pairs of keys indicated by the edges of G, resulting in $O(n^{2(1-\gamma_k)})=O(n^{\beta_k})$ comparisons. In the remaining $k-1$ rounds, apply the best algorithm (parallel decision tree) to merge, in $k-1$ rounds, all pairs of sets A_i, B_j (of size $f=\lfloor n^{\gamma_k}\rfloor$) for which A_iB_j is an edge of G^*. By Lemma 1 and the induction hypothesis, these $k-1$ rounds require at most

$$O(n^{1-\gamma_k})\cdot\text{MERGE}\,(k-1,n^{\gamma_k})=O(n^{1-\gamma_k}n^{\gamma_k\beta_{k-1}})=O(n^{\beta_k})$$

comparisons.

It turns out that we can establish a lower bound of the same order of magnitude. First we need the following fact:

LEMMA 2. *Let G be a bipartite graph with parts $\{v_1v_2,\cdots,v_n\},\{w_1,w_2,\cdots,w_n\}$ and let G have m edges. Then G contains a set of $\lceil m/2n\rceil$ edges of the form v_iw_{i+j} for some fixed j.*

Proof. Let $E_j = \{v_i w_{i'} \in E(G) | i' - i \equiv j \pmod n\}$, $E'_j = \{v_i w_{i'} \in E(G) | i' - i = j\}$ and $E''_j = \{v_i w_{i'} \in E(G) | n + i' - i = j\}$ for $j = 0, 1, \cdots, n-1$. Note that each E'_j and E''_j is a set of edges of the prescribed form $v_i w_{i+j}(v_i w_{i+(j-n)})$. Since $E(G) = \bigcup_{j=0}^{n-1} E_j$ and each $E_j = E'_j \cup E''_j$, we have partitioned $E(G)$ into $2n$ sets of edges $E'_0, E''_0, E'_1, E''_1, \cdots, E'_{n-1}, E''_{n-1}$. Hence one of the sets has at least $m/2n$ elements.

THEOREM 2. MERGE $(2, n) = \Omega(n^{4/3})$.

Proof. We shall show that MERGE $(2, n) \geqq \frac{1}{4} n^{4/3}$ for all large n. Otherwise, assume that there exists a parallel comparison tree to merge two subsets $A = \{a_1, \cdots, a_n\}$ and $B = \{b_1, \cdots, b_n\}$ of a linearly ordered set in two rounds with fewer than $\frac{1}{4} n^{4/3}$ comparisons. Then the (bipartite) undirected graph G, whose vertices are the keys and the edges the comparisons performed in the first round, has fewer than $\frac{1}{4} n^{4/3}$ edges. Let $f = \lfloor n^{1/3} \rfloor$ and $A_i = \{a_{(i-1)f+1}, a_{(i-1)f+2}, \cdots, a_{if}\}$, $B_i = \{b_{(i-1)f+1}, b_{(i-1)f+2}, \cdots, b_{if}\}$. Let $\tilde{G}$ be the (bipartite) undirected graph whose vertices are $A_1, A_2, \cdots, A_g, B_1, B_2, \cdots, B_g$ ($g = \lceil n/f \rceil$), and whose edges are all $A_i B_j$ such that G contains no edge joining any $a \in A_i$ to any $b \in B_j$. Then $\tilde{G}$ has more than $g^2 - \frac{1}{4} n^{4/3} \geqq \frac{3}{4} n^{4/3}$ edges and by Lemma 2, contains a set of edges $A_i B_{i+j}$ for a fixed j and $\frac{3}{4} n^{4/3}/2g > \frac{5}{16} n^{2/3}$ values of i.

Hence there are more than $\frac{5}{16} n^{2/3}$ sets A_i which have not had, in the first round, any element compared to any element of B_{i+j}. It is possible that in the unknown linear order each of the sets $A_i \cup B_{i+j}$ consists of consecutive elements. Then no arcs join the sets A_i and B_{i+j} in the extended transitive closure of the corresponding admissible orientation of G for any of the over $\frac{5}{16} n^{2/3}$ values of i. Hence, in the second round we would have to make more than

$$\tfrac{5}{16} n^{2/3} f^2 > \tfrac{1}{4} n^{4/3}$$

(for large n) comparisons. (The constant $\frac{1}{4}$ could by the same argument be replaced by any $c < \frac{1}{3}$.)

COROLLARY 2. *For each fixed* k, MERGE $(k, n) = \Omega(n^{\beta_k})$.

Proof. The argument is similar to that of Theorem 2, which is its initial step for an induction on $k \geqq 2$. (We have already observed that MERGE $(1, n) = n^2$.) We claim that for every $k \geqq 2$ there exists a constant c_k such that MERGE $(k, n) \geqq c_k n^{\beta_k}$ for large n. Assuming this holds for $k-1$, and setting $c_k = c_{k-1}/(2 + c_{k-1}) - \varepsilon$ (for any $\varepsilon > 0$), we shall show that the inequality holds for k. Otherwise, the graph G of the comparisons made in the first round has fewer than $c_k \cdot n^{\beta_k}$ edges. Letting $f = \lfloor n^{\gamma_k} \rfloor$ and defining $\tilde{G}$ as in the proof of Theorem 2, we find that $\tilde{G}$ contains more than $(2+\varepsilon)/(2 + c_{k-1}) n^{\beta_k}$ edges and by Lemma 2 more than $1/(2 + c_{k-1}) n^{\beta_k \gamma_k - 1}$ edges $A_i B_{i+j}$ for a fixed j. By the same argument as above, we may be required to merge $1/(2 + c_{k-1}) n^{\beta_k + \gamma_k - 1}$ pairs of sets of size f in the remaining $k-1$ rounds. By the induction hypothesis this will require at least

$$\left(\frac{1}{2 + c_{k-1}} n^{\beta_k + \gamma_k - 1}\right) c_{k-1} f^{\beta_{k-1}} > \left(\frac{c_{k-1}}{2 + c_{k-1}} - \varepsilon\right) n^{\beta_k + \gamma_k - 1 + \beta_{k-1}\gamma_k} = c_k n^{\beta_k}$$

comparisons in the worst case.

Thus the parallel comparison trees we have constructed to prove Corollary 1 are optimal, within a constant factor.

4. Sorting. We shall use the following strategy to sort n keys in k time intervals (i.e., by a parallel comparison tree of height k): Partition the n keys into $\lceil n^{\alpha} \rceil$ sets of sizes m and $m+1$ ($m \leqq n^{1-\alpha}$). Sort all sets in $k-j$ time intervals. In the remaining j time intervals merge together every pair of sets. (A dummy key may be used to make the sets of equal size.) We have proved:

LEMMA 3.

$$\text{SORT}(k, n) \leqq \lceil n^{\alpha} \rceil \cdot \text{SORT}(k-j, \lfloor n^{1-\alpha} \rfloor + 1) + \binom{\lceil n^{\alpha} \rceil}{2} \cdot \text{MERGE}(j, \lfloor n^{1-\alpha} \rfloor + 1)$$

for every α, $0 \leqq \alpha \leqq 1$ *and* j, $0 < j < k$.

By choosing a suitable α and applying Corollary 1, we shall obtain:

THEOREM 3. *For each fixed* k,

$$\text{SORT}(k, n) = O(n^{s_k}),$$

where $s_1 = 2$ *and* $s_k = \min (2(2^j - 1)s_{k-j} - 2^j)/((2^j - 1)s_{k-j} - 1)$ *with the minimum taken over all* j, $0 < j < k$, *for which* $s_{k-j} \geqq \beta_j$.

Let s_k be as defined above. We shall prove Theorem 3 by induction on k. Assuming $\text{SORT}(k-j, n) = O(n^{s_{k-j}})$ for $0 < j < k$ and applying Lemma 3, we obtain

$$\text{SORT}(k, n) = O(n^{\alpha} \cdot n^{(1-\alpha)s_{k-j}}) + O(n^{2\alpha} n^{(1-\alpha)\beta_j})$$

for any α and j, $0 \leqq \alpha \leqq 1$, $0 < j < k$. If $s_{k-j} \geqq \beta_j$ for some j, we let

$$\alpha = \frac{s_{k-j} - \beta_j}{1 + s_{k-j} - \beta_j}$$

and verify, from above that with this choice of α $\text{SORT}(k, n) = O(n^s)$, where

$$s = \frac{2s_{k-j} - \beta_j}{1 + s_{k-j} - \beta_j} = \frac{2(2^j - 1)s_{k-j} - 2^j}{(2^j - 1)s_{k-j} - 1}.$$

Since the choice of j was arbitrary (as long as $s_{k-j} \geqq \beta_j$), $\text{SORT}(k, n) = O(n^{s_k})$.

Note that Theorem 3 is not useful when $j = 1$. (It yields only $\text{SORT}(2, n) = O(n^2)$.) However, for other small values of k, it yields the following useful estimates:

$$\text{SORT}(3, n) = O(n^{8/5}),$$

$$\text{SORT}(4, n) = O(n^{20/13}),$$

$$\text{SORT}(5, n) = O(n^{28/19})$$

(observe that $\frac{28}{19} < \frac{3}{2}$). These bounds can be improved for $k \geqq 4$ by beginning the recurrence with, say, $s_2 = 1.667$ [4]; since this is at the expense of being actually able to construct the corresponding trees, we have not pursued the improvement.

COROLLARY 3. $\lim s_k = 1$.

Proof. We prove by induction on k that $s_{k+1} \leqq s_k$. Indeed, s_{k+1} is the minimum of terms ($j > 0$)

$$\frac{2(2^j - 1)s_{k+1-j} - 2^j}{(2^j - 1)s_{k+1-j} - 1}$$

for which $s_{k+1-j} \geqq \beta_j$. Each such term satisfies $s_{k-j} \geqq s_{k+1-j} \geqq \beta_j$ by the induction hypothesis and

$$\frac{2(2^j - 1)s_{k+1-j} - 2^j}{(2^j - 1)s_{k+1-j} - 1} \leqq \frac{2(2^j - 1)s_{k-j} - 2^j}{(2^j - 1)s_{k-j} - 1}$$

because the function $(2(2^j - 1)x - 2^j)/((2^j - 1)x - 1)$ is increasing on x. Hence each of the terms whose minimum is s_{k+1} is majorized by one of the terms whose minimum is s_k, with the sole exception of the last term of s_{k+1},

$$\frac{2(2^k - 1)s_1 - 2^k}{(2^k - 1)s_1 - 1} = \frac{3 \cdot 2^k - 4}{2 \cdot 2^k - 3} > \frac{3}{2}.$$

Therefore, for all relevant k the last term does not influence the computation of s_{k+1}, and hence $s_{k+1} \leqq s_k$. Furthermore, an even easier induction shows that $s_k \geqq 1$. Thus, the sequence s_k is monotone and bounded; let $L = \lim s_k$.

Define σ_j by the recurrence

$$\sigma_2 = 1, \qquad \sigma_j = 1 + \left(1 - \frac{1}{2^{j-1}-1}\right)\sigma_{j-1} \quad \text{for } j \geqq 3.$$

It is easily seen by induction on j, that $j/2 \leqq \sigma_j \leqq j-1$ for all $j \geqq 2$. These estimates shall turn out to be useful, as we are going to show that

$$s_{\binom{j}{2}} \leqq 1 + \frac{1}{\sigma_j}$$

for all $j \geqq 2$. For simplicity we shall write $J = \binom{j}{2}$ and $J' = \binom{j-1}{2}$. Recall that $\beta_i = 2^i/(2^i - 1)$, so that

$$s_J = \min_i \frac{2s_{J-i} - \beta_i}{1 + s_{J-i} - \beta_i}$$

where the minimum is taken over all i, $0 < i < J$, for which $s_{J-i} \geqq \beta_i$. We shall first prove, by induction of j, that $s_{\binom{j}{2}} \geqq \beta_j$ for all $j \geqq 2$. We assume that $s_{J'} \geqq \beta_{j-1}$ and proceed to show that $s_J \geqq \beta_j$. If $0 < i \leqq j$, then $s_{J-i} \geqq \beta_i$ implies $(2s_{J-i} - \beta_i)/(1 + s_{J-i} - \beta_i) \geqq \beta_i \geqq \beta_j$. If $j < i < J$, then

$$s_{J-i} \geqq s_{J'} \geqq \beta_{j-1} \geqq \frac{\beta_i + \beta_j - \beta_i\beta_j}{2 - \beta_j},$$

and hence $(2s_{J-i} - \beta_i)/(1 + s_{J-i} - \beta_i) \geqq \beta_j$ as well.

Next we prove $s_J \leqq 1 + 1/\sigma_j$ by induction on $j \geqq 2$. For that we choose $i = j - 1$ in the definition of s_J. Since $s_{J-(j-1)} = s_{J'} \geqq \beta_{j-1}$, we have

$$s_J \leqq \frac{2s_{J'} - \beta_{j-1}}{1 + s_{J'} - \beta_{j-1}} \leqq \frac{2(1 + 1/\sigma_{j-1}) - \beta_{j-1}}{1 + (1 + 1/\sigma_{j-1}) - \beta_{j-1}} = 1 + \frac{1}{\sigma_j}.$$

In conclusion, $s_{\binom{j}{2}} \leqq 1 + 2/j$, and $L = \lim s_k = 1$.

The inequality $s_{\binom{j}{2}} \leqq 1 + 1/\sigma_j$ can also be used to obtain estimates of SORT (k, n) for larger values of k. For instance, when $k = \binom{12}{2} = 66$ we have SORT $(66, n) = 0\,(n^{1.10\cdots})$ (which does not compare too unfavorably with SORT $(\infty, n) = \theta(n \log n)$, [7]). In a similar vein, the results of § 3 imply that

$$\text{MERGE}\,(10, n) = O(n^{1.001})$$

(while MERGE $(\infty, n) = \theta(n)$ [7]).

REFERENCES

[1] A. V. AHO, J. E. HOPCROFT AND J. D. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[2] B. BOLLOBÁS AND M. ROSENFELD, *Sorting in one round*, Israel J. Math., 38 (1981), pp. 154–160.

[3] D. R. FARRIS AND A. P. SAGE, *On the use of interpretive structural modeling for worth assessment*, Comput. and Electr. Eng., 2 (1975), pp. 149–174.

[4] R. HÄGGKVIST AND P. HELL, *Parallel sorting with constant time for comparisons*, SIAM J. Comput., 10 (1981), pp. 465–472.

[5] ———, *Graphs and parallel comparison algorithms*, Proc. XI Southeastern Conference on Combinatorics, Graph Theory and Computing, Congressus Numerantium, 29 (1980), pp. 497–509.

[6] W. L. HART AND D. W. MALONE, *Goal setting for a state environmental agency*, Proc. 1974 IEEE Conference on Decision and Control, Phoenix, Arizona, 1974.
[7] E. HOROWTIZ AND S. SAHNI, *Fundamentals of Computer Algorithms*, Computer Science Press, Potomac, MD, 1978.
[8] D. KNUTH, *The Art of Computer Programming, Vol. 3, Sorting and Searching*, Addison-Wesley, Reading, MA, 1973.
[9] F. S. ROBERTS, *Measurement Theory*, Encyclopedia of Mathematics and its Applications, vol. 7, Addison-Wesley, Reading, MA, 1979.
[10] S. SCHEELE, *Final report to U.S. Office of Environmental Education, Department of Health, Education and Welfare, Project to develop shared conceptual tools for environmental education*, Social Engineering Technology, Los Angeles, 1977.
[11] L. VALIANT, *Parallelism in comparison problems*, SIAM J. Comput., 4 (1975), pp. 348–355.
[12] R. J. WALLER, *Applications of interpretive structural modeling in management of the learning disabled*, in Portraits of Complexity, M. M. Baldwin, ed., Battelle Monograph No. 9, Battelle Memorial Institute, Columbus, OH, 1975, pp. 95–103.
[13] ———, *An application of interpretive structural modeling to priority setting in urban systems management*, ibid., pp. 104–108.
[14] J. WARFIELD, *Binary matrices in system modeling*, IEEE Trans. Systems, Management and Cybernetics, SMC-3 (1973), pp. 441–449.
[15] ———, *Developing subsystem matrices in structural modeling*, IEEE Trans. Systems, Management and Cybernetics, SMC-4 (1974), pp. 74–80.
[16] ———, *Developing interconnection matrices in structural modeling*, IBM Trans. Systems, Management and Cybernetics, SMC-4 (1974), pp. 81–87.
[17] ———, *Societal Systems*, Wiley-Interscience, New York, 1976.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0008 $01.00/0

THE BOUNDED PATH TREE PROBLEM*

PAOLO M. CAMERINI† AND GIULIA GALBIATI‡

Abstract. The subject of this paper is the bounded path tree (BPT) problem: An undirected graph $G=(V,E)$ is given whose edges have nonnegative lengths; two subsets I and J of V are also given, and nonnegative constants U_i, W_j are associated with each $i\in I$, $j\in J$. The BPT problem asks for a tree of G whose vertex set contains $I\cup J$ and whose path joining vertices i and j is not longer than U_i+W_j, for each $i\in I$, $j\in J$. This problem generalizes the shortest path and the minimum longest path spanning tree problem. It complements standard min–max location problems, as it asks for a tree given the facility locations, instead of locating facilities in a given network. In this paper we propose some applications of the BPT problem for the design of emergency and communication networks, show its equivalence to an extension of the absolute center location problem and give an algorithm for its solution. This algorithm requires time $O(k|E|+k|V|\log k)$, where $k=|I\cup J|$, plus time for finding in G all shortest path lengths between a vertex in $I\cup J$ and a vertex in V. We also consider a few simple extensions of the BPT problem, such as those admitting negative or multiple edge lengths, lower (as well as upper) bounds to path lengths, constants Z_{ij} instead of U_i+W_j. We show that all these extensions are NP-complete.

1. Problem statement and presentation. In this paper we study the problem of finding trees with bounded path lengths between pairs of vertices. Specifically, we consider the *bounded path tree* (BPT) *problem*, which can be stated as follows.

An undirected graph $G=(V,E)$ is given, where $V=\{1,\cdots,n\}$ is the set of *vertices* and $E\subseteq\{\{i,j\}\,|\,i,j\in V, i\neq j\}$ is the set of *edges*, $|E|=m$. A nonnegative, real-valued function $w:E\to\mathbb{R}^+$ is associated with G, and for each $e\in E$, $w(e)$ is called the *length* of e. Two subsets I, J of V are also given, and nonnegative real numbers U_i, W_j are associated with each $i\in I$, $j\in J$.

The BPT problem asks for a tree T of G, such that

(a) the vertex set of T contains $I\cup J$;

(b) $\lambda(i,j,T)\leqq U_i+W_j$ for each $i\in I$, $j\in J$.

Here and in what follows, $\lambda(i,j,T)$ denotes the *length* of the path $\pi(i,j,T)$, i.e., the sum of the edge lengths in the unique path of T joining vertices i and j. Conventionally, $\lambda(i,j,T)=0$ when $i=j$.

Any tree T satisfying conditions (a) and (b) above is called a *bounded path tree*.

Notice that when $|I|=1$ and $J=V$, the above problem is equivalent to the shortest path tree problem [4]. When $I=J=V$ and the bound U_i+W_j is independent of i and j, the BPT problem is equivalent to the minimum longest path spanning tree problem, whose complexity and close relationship with the absolute center location problem have been discussed in [1], [7]. We shall see in §§ 3 and 4 that a similar relationship exists between the general BPT problem and an extended version of the absolute center location problem.

2. Applications. The BPT problem has many natural applications, for instance, in the design of emergency networks. Here edge lengths represent travel times between centers; I is the set of centers which may require service for emergency, J is the set of centers where emergency facilities are located. The constants U_i's and W_j's can model priority levels and/or local service times corresponding to demand centers and to facilities, respectively. As discussed in [7, p. 87–88], additive rather than multiplicative factors are often appropriate in modeling such kind of situations.

* Received by the editors September 9, 1981.

† Centro di Studio per le Telecomunicazioni Spaziali, Consiglio Nazionale delle Ricerche, Politecnico di Milano, Piazza L. da Vinci, 32, 20133 Milano, Italy.

‡ Istituto di Matematica, Università di Pavia, 27100 Pavia, Italy.

In this context the BPT problem complements classical min–max location problems since the former asks for finding a tree given the demand centers and the facility locations, whereas the latter ask for locating facilities given the network and the demand centers.

Further applications of the BPT problem derive naturally from its equivalence to an extension of the absolute center location problem, which is studied in § 3.

Moreover, the BPT problem can be applied to the design of reliable communication networks. Here for each edge $e=\{u, v\}$ (vertex v) of G, let $p_{uv}(p_v)$ be a given probability for the edge (vertex) to be "alive". Assuming independence, the probability for a path to be "alive" is the product of the probabilities for its edges and vertices to be "alive". If we take

$$w(\{u, v\}) = -\log(\sqrt{p_u} \cdot p_{uv} \cdot \sqrt{p_v})$$

for each $e=\{u, v\}$, then for any tree T

$$-\lambda(i, j, T) + \log \sqrt{p_i} + \log \sqrt{p_j}$$

is the logarithm of the probability to be "alive" for the path $\pi(i, j, T)$ joining in T vertices i and j.

Assume now $I=J=V$ and take for all $i \in V$

$$U_i = W_i = \log \sqrt{p_i} - \tfrac{1}{2} \log A,$$

A being a given constant, $0 \leqq A \leqq 1$.

In this case the BPT problem asks for a tree T such that the probability for each path $\pi(i, j, T)$ to be "alive" is not less than A.

3. Notation and discussion. In this section we introduce some notation and discuss some connections between the BPT problem and the absolute center location problem in more detail than in § 1. As a consequence of this discussion, the algorithm proposed in § 4 for solving the BPT problem may be better understood and viewed as an extension of Hakimi's method [6] for the absolute center location problem.

In order to formalize these concepts, we first recall some terminology, related to the notion of "points" and "lines" of an undirected, edge weighted graph. Referring to the graph G and the weighting function w of § 1, we define a *point* of G to be either a vertex or an ordered pair (e, θ), where e is an edge with positive length and θ is a real number such that $0<\theta<w(e)$. In this latter case, the point is called an *internal* point of e. We agree that edges with zero length have no internal points, and the vertices i and j of any edge e are called *end* points of e. Assuming without loss of generality that $i<j$, we shall also write $(e, 0)$ for i and $(e, w(e))$ for j.

In the usual geometrical representation Γ of the graph G, the points of G are represented by geometric points of Γ. As an example Fig. 1 illustrates the geometrical representation of a point $x=(e, \theta)$, with $0 \leqq \theta \leqq w(e)$, $w(e)>0$.

As the notion of vertices is extended to that of points, similarly the notion of edges can be extended to that of lines, by saying that if $x=(e, \theta)$, $x'=(e, \theta')$ are two (not necessarily distinct) points of an edge $e \in E$, then the unordered pair $l=\{x, x'\}$ is a *line* of G. The function w can also be extended to the set of lines, by defining $w(l)=|\theta-\theta'|$ to be the *length* of line l.

Accordingly, the definition of path joining two vertices is extended to define a *path (of G) joining two points* x and y. Such a path is a set P of $p \geqq 1$ distinct lines of G of the form

$$P = \{\{x=x_0, x_1\}, \{x_1, x_2\}, \cdots, \{x_{p-1}, x_p=y\}\},$$

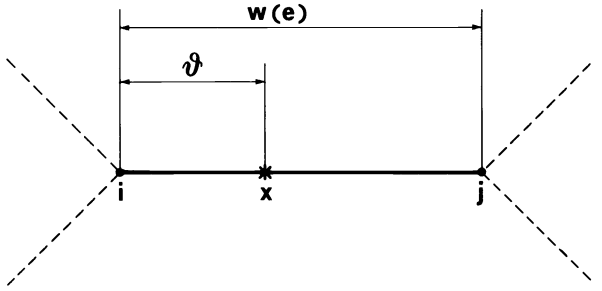


FIG. 1. *Geometrical representation of a point x of edge* $e = \{i, j\}$, *with* $i < j$, $w(e) > 0$.

where $x_1, x_2, \cdots, x_{p-1}$ must be distinct vertices of G, all different from x and y. When $x = y$, the path is *closed*. The *length* of P is given by

$$\sum_{h=1}^{p} w(\{x_{h-1}, x_h\}).$$

A closed path with more than one line is a *cycle*.

As usual we call a *tree* a connected, acyclic subgraph of G, i.e., a subgraph T of G having exactly one path joining any two points x, y of T. According to what was stated in § 1, such a path is denoted by $\pi(x, y, T)$, and $\lambda(x, y, T)$ indicates its length. An *absolute center* of a tree T is a point in the middle of a longest path of T, i.e., a point z of some edge of $\pi(s, t, T)$, such that

$$\lambda(s, z, T) = \lambda(z, t, T),$$

and $\pi(s, t, T)$ is a longest path joining two vertices of T.

Assume now that G is connected. Since line lengths are nonnegative, there always exists a shortest path joining any two points x and y. The length of this path is called *distance* between x and y and is denoted by $\delta(x, y)$. Given any point z and any subset S of V, there always exists a set of shortest paths joining z and each vertex of S such that the set L of their lines does not contain cycles. This set identifies a *tree of shortest paths originating* at z and *terminating at S*. This tree is denoted by $T(z, S)$ and its edge set is made up of all edges in L, with the addition of edge $e = \{i, j\}$ if z is an internal point of e and L contains both lines $\{z, i\}$ and $\{z, j\}$.

Referrring to the above terminology, consider the following two problems where W is a given nonnegative real number and G is a connected graph.

Problem 1 (Spanning tree with bounded longest path). Find a spanning tree T of G such that $\lambda(i, j, T) \leqq 2W$ for all $i, j \in V$.

Problem 2 (Absolute center location). Find a point z of G such that $\delta(z, i) \leqq W$ for all $i \in V$.

It has been shown [1], [7, pp. 113–114] that these two problems are equivalent. Specifically, if z is a solution to Problem 2, then any tree $T(z, V)$ is a solution to Problem 1, while the absolute center z of a tree T, solution to Problem 1, is a solution to Problem 2.

Hakimi's algorithm [6] may be viewed as a method for solving both problems. Its efficiency is mainly due to the fact that the search for a solution to Problem 2 may be restricted to the vertices and to those internal points of G which are local minima of the function

$$\gamma(x) = \max_{v \in V} \delta(x, v),$$

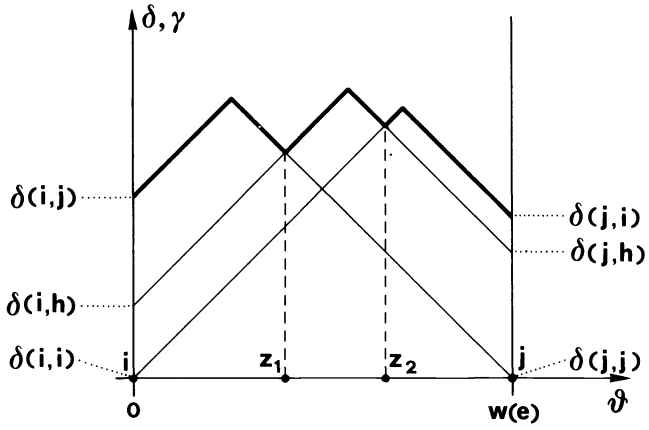


FIG. 2. *Representation of* $\gamma(x)$ *and* $\delta(x, v)$ $(v = h, i, j)$, *restricted to the points* $x = (e, \theta)$ *of edge* $e = \{i, j\}$, $i < j$, $w(e) > 0$. *The graph of* $\gamma(x)$ *(drawn in heavy lines) is the upper envelope of* $\delta(x, v)$, $v = h, i, j$.

defined over the set of points of G. Examples of such local minima are points z_1 and z_2 in Fig. 2. Since each function $\delta(x, v)$ is piece-wise linear with slope ± 1, local minima of $\gamma(x)$ can be easily computed from the distances between all pairs of vertices.

These ideas can be extended to the BPT and to the following problem, where I, J, U_i, $W_j (i \in I, j \in J)$ are given as in § 1.

Problem 3 (Bounded path absolute center). Find a point z of G, such that $\delta(z, i) + \delta(z, j) \leqq U_i + W_j$ for each $i \in I$, $j \in J$.

Referring to the interpretation given in § 2, the bounded path absolute center (BPC) problem may be viewed as that of locating a "center", which must be visited along every path from a facility location $(j \in J)$ to a demand vertex $(i \in I)$.

The following theorem states the equivalence between the BPT and the BPC problems.

THEOREM 1. *If* z *is a solution to the* BPC *problem, then any tree of shortest paths originating at* z *and terminating at* $I \cup J$ *is a* BPT *of* G. *If* T *is a* BPT *of* G, *then at least one point of* T *is a solution to the* BPC *problem.*

Proof. Let $\bar{T} = T(z, I \cup J)$, where z is a solution to the BPC problem. For each $i \in I$, $j \in J$, we have

$$\lambda(i, j, \bar{T}) \leqq \delta(z, i) + \delta(z, j) \leqq U_i + W_j,$$

so that $\bar{T}$ is a BPT of G, and the first part of the theorem is proved.

Let now T be a BPT of G, and let

$$R = \min_{x,y \in I} \{U_x + U_y - \lambda(x, y, T)\} = U_s + U_t - \lambda(s, t, T).$$

There are two cases.

Case 1. For some $v \in I$, $R \geqq 2U_v$. Then

$$U_i - \lambda(v, i, T) \geqq U_v \quad \text{for each } i \in I. \tag{1}$$

Since T is a BPT of G,

$$W_j - \lambda(v, j, T) \geqq -U_v \quad \text{for each } j \in J. \tag{2}$$

From (1) and (2) it follows that

$$\delta(v, i) + \delta(v, j) \leqq \lambda(v, i, T) + \lambda(v, j, T) \leqq U_i + W_j$$

for each $i \in I$, $j \in J$; i.e., v is a point of T, solution to the BPC problem.

Case 2. For each $v \in I$, $R < 2U_v$. Then

$$|U_s - U_t| < \lambda(s, t, T),$$

and there exists a point $\bar{z}$ of some edge e of $\pi(s, t, T)$ such that

$$U_s - \lambda(s, \bar{z}, T) = U_t - \lambda(\bar{z}, t, T) = \frac{R}{2}.$$

We can now show that

$$U_i - \lambda(i, \bar{z}, T) \geqq \frac{R}{2} \quad \text{for each } i \in I \tag{3}$$

and

$$W_j - \lambda(j, \bar{z}, T) \geqq -\frac{R}{2} \quad \text{for each } j \in J. \tag{4}$$

In order to prove these inequalities, assume without loss of generality that for each $i \in I$, $j \in J$, vertices s, i and j belong to the same tree when e is removed from T; otherwise, interchange appropriately the roles of s and t in what follows. Since

$$\lambda(i, t, T) = \lambda(i, \bar{z}, T) + \lambda(\bar{z}, t, T), \qquad \lambda(s, t, T) = \lambda(s, \bar{z}, T) + \lambda(\bar{z}, t, T)$$

and by definition of R

$$U_i + U_t - \lambda(i, t, T) \geqq U_s + U_t - \lambda(s, t, T),$$

we have that

$$U_i - \lambda(i, \bar{z}, T) \geqq U_s - \lambda(s, \bar{z}, T) = \frac{R}{2}$$

for each $i \in I$, and inequality (3) is proved.

Since

$$\lambda(j, t, T) = \lambda(j, \bar{z}, T) + \lambda(\bar{z}, t, T),$$

$$\lambda(\bar{z}, t, T) = U_t - \frac{R}{2}$$

and

$$\lambda(j, t, T) \leqq W_j + U_t,$$

we have that

$$W_j - \lambda(j, \bar{z}, T) \geqq W_j - \lambda(j, t, T) + U_t - \frac{R}{2} \geqq -\frac{R}{2}$$

for each $j \in J$, and inequality (4) is proved.

Because of (3) and (4), we may write

$$\delta(\bar{z}, i) + \delta(\bar{z}, j) \leqq \lambda(\bar{z}, i, T) + \lambda(\bar{z}, j, T) \leqq U_i + W_j$$

for each $i \in I$, $j \in J$, and in both Cases 1 and 2, we may conclude that there exists a point of T, solution to the BPC problem. □

In the next section, we present an algorithm for solving both the BPT and the BPC problems. In a similar way as in Hakimi's algorithm, the search for a solution

to the BPC problem is restricted to the vertices and to the *peaks* of ϕ, i.e., the internal points of G, which are local maxima of the function

$$\phi(x) = \min_{i \in I} \{U_i - \delta(x, i)\}.$$

Alternatively, the search could be carried out among the vertices and the peaks of

$$\psi(x) = \min_{j \in J} \{W_j - \delta(x, j)\}.$$

4. Algorithm. In this section we describe an algorithm for solving both the BPT and the BPC problems and evaluate its complexity.

The following procedure receives as input the items G, W, I, J, U_i, W_j $(i \in I, j \in J)$ and produces the output 'yes' or 'no', depending upon whether or not the graph G contains a solution for both problems. For the sake of simplicity, we assume that G is connected, so that functions ϕ and ψ are defined at all points of G. The comment of step 3 refers to the implementation of step 5, described at the end of this section, for analyzing its complexity.

procedure BOUNDEDPATH:
begin
1. **for** each $u \in I \cup J$, $v \in V$ **do** compute $\delta(u, v)$;
2. **for** each $u \in V$ **do** form two lists, containing the vertices i of I (respectively, j of J) in nondecreasing order of $U_i - \delta(u, i)$ (respectively, $W_j - \delta(u, j)$);
3. **comment** steps 1 and 2 above allow an efficient implementation of step 5;
4. **for** each $z \in V$ **do if** $\phi(z) + \psi(z) \geqq 0$ **then return** 'yes';
5. **for** each peak z of ϕ **do if** $\phi(z) + \psi(z) \geqq 0$ **then return** 'yes';
6. **return** 'no'

end

The following theorem proves the correctness of this procedure and shows how to utilize it for finding whenever it exists a solution to either the BPT or the BPC problem.

THEOREM 2. *If* BOUNDEDPATH *returns 'yes', then the current point z and the tree $T(z, I \cup J)$ are solutions to the* BPC *and the* BPT *problem, respectively. If* BOUNDEDPATH *returns 'no' then neither problem has a solution.*

Proof. If BOUNDEDPATH returns 'yes', then the current point z is such that $\phi(z) + \psi(z) \geqq 0$, i.e.,

$$\delta(z, i) + \delta(z, j) \leqq U_i + W_j,$$

for each $i \in I, j \in J$. Thus, z is a solution to the BPC problem. By Theorem 1, $T(z, I \cup J)$ is a solution to the BPT problem, and the first part of Theorem 2 is proved.

Consider now the case where BOUNDEDPATH returns 'no'. Suppose that there exists a point z of G, solution to the BPC problem. For this point $\phi(z) + \psi(z) \geqq 0$. If z is a vertex or a peak of ϕ, we have a contradiction since the procedure would have returned 'yes'. If z is an internal point of $e = \{i, j\}$ and is not a peak of ϕ, then moving from z, either towards i or towards j on edge e, makes the value of ϕ increase. Let us move in such a direction until either a vertex or a peak of ϕ is found. For this point z, $\phi(z') > \phi(z)$.

Due to the special form of the functions ϕ and ψ, it is easy to see that if we move from z to z', the value of ψ can not decrease more than the value of ϕ increases, so that $\phi(z') + \psi(z') \geqq 0$. This again contradicts the fact that BOUNDEDPATH returns

'no'. Thus, neither the BPC nor (because of Theorem 1) the BPT problem has a solution, and Theorem 2 is proved. □

We now examine the time complexity of BOUNDEDPATH.

Step 1 requires to compute all distances between a vertex in $I \cup J$ and a vertex in V. This task can be performed by many different algorithms [8], [3], the best choice depending on the range of the edge lengths, the sparseness of the graph and so on. Thus, we do not specify the time complexity of step 1 and only remind to include it in the overall time complexity evaluation. We remark that the identification of the shortest paths is not required in step 1, so that computing distances without actually identifying the corresponding shortest paths might be preferable. From the time complexity viewpoint, however, it is still unknown whether distances can be computed faster than shortest paths [9], [10].

Step 2 sorts $n = |V|$ times the two sets I and J, so that its time complexity is $O(kn \log k)$, where $k = |I \cup J|$.

Using the first vertices in the two lists obtained in step 2 for vertex z, each condition $\phi(z) + \psi(z) \geqq 0$ in step 4 can be tested in $O(1)$ time, and hence, step 4 needs time $O(n)$.

Finally, step 5 can be implemented to use $O(k \cdot |E|)$ time: For each edge $e = \{u, v\} \in E(u < v)$ such that $\bar{w} = w(e) > 0$, we do the following.

First, we extract from I a sequence $i_1, \cdots, i_p$, called *irredundant*, having the following properties.

1) Either $p = 1$ or $p > 1$, in which case $a_h < a_{h+1}$ and $b_h > b_{h+1}$ for $h = 1, \cdots, p-1$, where $a_h = U_{i_h} - \delta(u, i_h)$ and $b_h = U_{i_h} - \delta(v, i_h)$ for $h = 1, \cdots, p$.

2) $\phi(x) = \min_{h=1,\cdots,p} \{U_{i_h} - \delta(x, i_h)\}$ for all points x of e.

The example of Fig. 3 illustrates these properties and should convince the reader—without formal proof—that an irredundant sequence always exists. It is also easy to see that such a sequence can be found by scanning only once each vertex i of I in nondecreasing order of $U_i - \delta(u, i)$ (or $U_i - \delta(v, i)$). Since this ordering has already been made in step 2, the time needed for finding the irredundant sequence is $O(|I|)$.

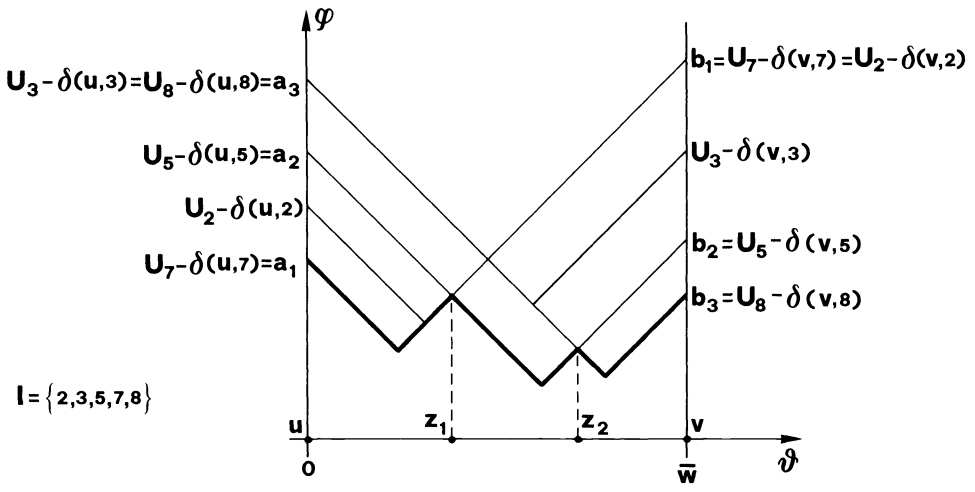


FIG. 3. *Representation of $\phi(x)$ restricted to the points $x = (e, \theta)$ of edge $e = \{u, v\}$, $u < v$, $\bar{w} = w(e) > 0$. The graph of $\phi(x)$ is drawn in heavy lines. The sequence 7, 5, 8 is irredundant.*

Second, the (at most) $2p$ linear intervals of the piecewise linear function $\phi(x)$ —restricted to the points $x = (e, \theta)$—are identified by the following sequence of values of θ:

$$0, \quad \tfrac{1}{2}(a_1 - b_1 + \bar{w}), \quad \tfrac{1}{2}(a_2 - b_1 + \bar{w}), \quad \tfrac{1}{2}(a_2 - b_2 + \bar{w}), \quad \cdots, \quad \tfrac{1}{2}(a_p - b_{p-1} + \bar{w}), \quad \tfrac{1}{2}(a_p - b_p + \bar{w}), \quad \bar{w}. \tag{5}$$

(The first two and/or the last two values may coincide.) In particular, if $p > 1$ the $p - 1$ peaks of edge e and the corresponding values of ϕ are given by $z_h = (e, \theta_h)$, where

$$\theta_h = \tfrac{1}{2}(a_{h+1} - b_h + \bar{w})$$

and

$$\phi(z_h) = \tfrac{1}{2}(a_{h+1} + b_h - \bar{w})$$

for $h = 1, \cdots, p - 1$. These values can be obtained in $O(|I|)$ time.

Third, if $p > 1$, we determine in $O(|J|)$ time the linear intervals of ψ by repeating for ψ the computations made above for ϕ. The $p - 1$ values of ψ at the peaks of ϕ can be obtained from the linear intervals of ψ, by means of linear interpolations. The time required is $O(|I| + |J|)$, i.e., $O(k)$.

We dispense with giving a formal procedure implemented along the above lines. However it should be clear that such a procedure allows executing step 5 of BOUNDEDPATH in $O(km)$ time, where $m = |E|$.

The time complexity of BOUNDEDPATH is therefore $O(km + kn \log k)$, plus time for the computations of step 1.

We conclude this section with some practical remarks.

Remark 1. When G is even moderately dense, e.g., m is $\Omega(n^{1+\varepsilon})$ for some $\varepsilon > 0$, the term km dominates $kn \log k$ in the time complexity of BOUNDEDPATH.

Remark 2. If a vertex cover of size c is available, i.e., a set of $c \leqq n$ vertices of G shares at least one vertex with each edge of G, the term $kn \log k$ in the time complexity of BOUNDEDPATH may be reduced to $kc \log k$.

Remark 3. By Theorem 2 the time needed to solve the BPC problem is the same as that required by BOUNDEDPATH.

Remark 4. By Theorem 2 the time needed to solve the BPT problem is that required by BOUNDEDPATH, plus the time for finding $T(z, I \cup J)$ in case of a 'yes' answer. This extra time, however, is $O(km)$ and, hence, is dominated by the time complexity of BOUNDEDPATH. In fact, assume that z is an internal point of $e = \{u, v\}$. For each vertex h of $I \cup J$, a shortest path joining h and u (or v) can be obtained in $O(m)$ time by backtracking from u (or v): recall that all distances between vertex h and the vertices of V have been computed in step 1 of BOUNDEDPATH. Moreover, if all shortest paths corresponding to distances have already been identified in step 1, the time needed for finding $T(z, I \cup J)$ is only $O(n)$.

Remark 5. Bounding techniques similar to those suggested in [7, pp. 120–125] or [2, pp. 90–105] can be helpful for restricting the search in step 5 of BOUNDEDPATH to a (hopefully small) subset of the edge set E.

Remark 6. When the edge lengths are uniform, i.e., $w(e) = 1$ for each $e \in E$, and U_i, W_j are nonnegative integers for each $i \in I$, $j \in J$, then from (5) it is easy to see that any peak of either ϕ or ψ must be a point $z = (e, 1/2)$ for some edge e. It follows that step 5 of BOUNDEDPATH can be executed in time $O(km)$, since for each edge e, the values $\phi((e, 1/2))$ and $\psi((e, 1/2))$ can be obtained in $O(|I|)$ and $O(|J|)$ time, respectively, without using the lists computed in step 2. Therefore, the whole com-

plexity of BOUNDEDPATH becomes $O(k \cdot m)$, since in this case step 1 needs $O(k \cdot m)$ time.

5. Extensions and NP-completeness. In this section we focus our attention on the fact that even very simple extensions or modifications of the BPT problem lead to NP-complete decision problems [5], which are therefore probably not solvable with algorithms having polynomial time complexity.

The first extension that we consider is the one admitting negative edge lengths and negative real numbers U_i and W_j associated with some vertices $i \in I$ and $j \in J$. In order to show that this extended problem is NP-complete, it is enough to notice that when the input is a graph G with $n > 1$, $w(e) = -1$ for each $e \in E$, $I = \{1\}$, $J = \{n\}$ and $U_1 = W_n = -(n-1)/2$, we are dealing with the problem of detecting in G the existence of an Hamiltonian path joining vertices 1 and n, and this problem is known to be NP-complete [5, p. 60].

A similar reasoning can be used to show that the analogue of the BPT problem, where we require the quantity $U_i + W_j$ to bound $\lambda(i, j, T)$ from below rather than from above, is also NP-complete. In fact, this problem, when restricted to graphs G having $n > 1$, $w(e) = 1$ for each $e \in E$, $I = \{1\}$, $J = \{n\}$, $U_1 = W_n = (n-1)/2$, again becomes the problem of detecting in G an Hamiltonian path joining vertices 1 and n.

Let us now consider another extension of the BPT problem. For each $i \in I$, $j \in J$ a nonnegative real number Z_{ij} is given instead of two nonnegative real numbers U_i, W_j, and the problem asks for a tree T of G such that

(a) the vertex set of T contains $I \cup J$;

(b') $\lambda(i, j, T) \leqq Z_{ij}$ for each $i \in I$, $j \in J$.

In order to prove the NP-completeness of this extended problem, we exhibit a polynomial time transformation to it from the NP-complete SATISFIABILITY problem (SAT) [5, p. 39].

In other terms, for every input of SAT, we define a corresponding input of our problem that can be computed in polynomial time, and we show that answers to the corresponding inputs are either both 'yes' or both 'no'. For convenience of the reader, we recall that SAT can be stated as follows.

INPUT. A family $C = \{c_1, \cdots, c_q\}$ of clauses over a finite set $L = \{x_1, \cdots, x_h, \bar{x}_1, \cdots, \bar{x}_h\}$ of literals.

QUESTION. Is there a truth assignment for L that satisfies all the clauses of C?

Let an input of SAT be given as above. Define a corresponding input of our problem as follows:

$$V = \{\rho\} \cup L \cup C,$$

$$E = \bigcup_{r=1}^{h} \{\{\rho, x_r\}, \{\rho, \bar{x}_r\}, \{x_r, \bar{x}_r\}\}$$

$$\cup \{\{x_r, c_s\} \mid x_r \in c_s, r \in \{1, \cdots, h\}, s \in \{1, \cdots, q\}\}$$

$$\cup \{\{\bar{x}_r, c_s\} \mid \bar{x}_r \in c_s, r \in \{1, \cdots, h\}, s \in \{1, \cdots, q\}\},$$

$$I = J = V,$$

$$w(e) = 1 \quad \text{for each } e \in E,$$

$$Z_{ij} = \begin{cases} 1 & \text{if } \{i, j\} = \{x_r, \bar{x}_r\} \quad \text{for some } r \in \{1, \cdots, h\}, \\ 2 & \text{if } \{i, j\} = \{\rho, c_s\} \quad \text{for some } s \in \{1, \cdots, q\}, \\ \infty & \text{otherwise.} \end{cases}$$

We claim that there is a truth assignment for L that satisfies all the clauses of C if and only if $G=(V,E)$ contains a tree T that satisfies (a) and (b'). In fact, if there exists a truth assignment that satisfies the clauses, denote by L' the subset of L containing the literals which have been assigned the value 'true'. Of course, L' does not contain a complementary pair of literals. For each $s=1,\cdots,q$, choose in L' a literal x_t (or $\bar{x}_t$) that satisfies c_s and consider the three edges $\{\rho, x_t\}$, $\{x_t, c_s\}$ (or $\{\rho, \bar{x}_t\}$, $\{\bar{x}_t, c_s\}$) and $\{x_t, \bar{x}_t\}$. The union of these edges forms a tree T of G which can be made into a spanning tree by adding edges $\{\rho, x_r\}$, $\{x_r, \bar{x}_r\}$, if x_r is not already a vertex of T, for each $r=1,\cdots,h$. The resulting tree satisfies (a) and (b'). Conversely, assume G has a tree T for which (a) and (b') hold. We can deduce the following facts.

(i) Since $\lambda(\rho, c_s, T)\leqq 2$ for each $s=1,\cdots,q$, every path in T joining ρ and a clause has length two and has only one intermediate vertex.

(ii) If L' is the set of the intermediate vertices on the paths joining ρ and all the clauses, then L' cannot contain both a literal and its complement since $\lambda(x_r, \bar{x}_r, T)\leqq 1$, $r=1,\cdots,h$.

Therefore, if we assign the value 'true' to the literals in L', all the clauses are satisfied, i.e., there exists a truth assignment for L satisfying all the clauses of C.

Finally, consider the extension of the BPT problem obtained when to the graph G are associated two nonnegative real-valued functions, w_1 and w_2 instead of one, and the problem asks for a tree T of G such that

(a) the vertex set of T contains $I\cup J$;

(b'') $\lambda_1(i,j,T)\leqq U_i+W_j$, $\lambda_2(i,j,T)\leqq U_i+W_j$ for each $i\in I$, $j\in J$, where $\lambda_1(i,j,T)$ and $\lambda_2(i,j,T)$ denote the lengths of $\pi(i,j,T)$, corresponding to w_1 and w_2, respectively.

In order to prove that this problem is NP-complete, we use a transformation from SAT to our problem similar to the one defined above. For any input of SAT, the corresponding input of this problem is given by the graph $G=(V,E)$ defined above and by the following items.

$I=\{\rho\}$,

$J=C$,

$$w_1(e)=\begin{cases}1 & \text{if } e=\{\rho,x_r\} \text{ or } e=\{x_r,c_s\} \text{ for some } s\in\{1,\cdots,q\} \text{ and } r\in\{1,\cdots,h\},\\ 0 & \text{otherwise,}\end{cases}$$

$$w_2(e)=\begin{cases}1 & \text{if } e=\{\rho,\bar{x}_r\} \text{ or } e=\{\bar{x}_r,c_s\} \text{ for some } s\in\{1,\cdots,q\} \text{ and } r\in\{1,\cdots,h\},\\ 0 & \text{otherwise,}\end{cases}$$

$U_\rho=1$, $\quad W_{c_s}=0$ for each $s=1,\cdots,q$.

Again we claim that all the clauses of C can be satisfied by a truth assignment for L if and only if G contains a tree satisfying (a) and (b'').

If there exists a truth assignment that satisfies the clauses, a tree T satisfying (a) and (b'') can be constructed in the same way as in the preceding transformation but for the fact that here we consider the three edges $\{\rho, \bar{x}_t\}$, $\{x_t, c_s\}$, $\{x_t, \bar{x}_t\}$ or $\{\rho, x_t\}$, $\{\bar{x}_t, c_s\}$, $\{x_t, \bar{x}_t\}$, depending on the literal chosen being x_t or $\bar{x}_t$.

Conversely, if G has a tree T satisfying (a) and (b''), then for each $s=1,\cdots,q$ the path $\pi(\rho, c_s, T)$ is either $\{\{\rho, \bar{x}_t\}, \{\bar{x}_t, x_t\}, \{x_t, c_s\}\}$ or $\{\{\rho, x_t\}, \{x_t, \bar{x}_t\}, \{\bar{x}_t, c_s\}\}$ for some $t\in\{1,\cdots,h\}$. Therefore, if we take x_t or $\bar{x}_t$—respectively—for each clause c_s, we obtain a set L' of literals, no two of which are complementary. Similarly as in the preceding transformation, it follows that a truth assignment for L can be constructed which satisfies all the clauses of C.

6. Conclusions and remarks. We have proposed an algorithm for solving in polynomial time a couple of equivalent problems—the bounded path tree and the bounded path absolute center problem. These problems generalize two other known, polynomially solvable equivalent problems—the minimum longest path spanning tree and the absolute center location problem. The proposed algorithm reflects this generalization and can be easily modified—maintaining the same time complexity—to treat a slightly more general formulation of these problems, in which some (or all) the path bounding constraints are substituted with a single min–max objective function. We leave to the interested reader the task of elaborating by himself such an extension as we feel that the formal description of an extended algorithm would be more cumbersome than theoretically relevant.

We conclude by remarking that the BPT problem is, in a sense, a "most difficult easy" problem, as we have shown that even simple generalizations of it lead to NP-complete decision problems.

Acknowledgment. We are very grateful to our friend Francesco Maffioli for his suggestions and encouragements.

REFERENCES

[1] P. M. CAMERINI, G. GALBIATI AND F. MAFFIOLI, *Complexity of spanning tree problems: Part* I, European J. Operational Research, 5 (1980), pp. 346–352.

[2] N. CHRISTOFIDES, *Graph Theory: An Algorithmic Approach*, Academic Press, New York, 1975.

[3] R. DIAL, F. GLOVER, D. KARNEY AND D. KLINGMAN, *A computational analysis of alternative algorithms and labeling techniques for finding shortest path trees*, Networks, 9 (1979), pp. 215–248.

[4] E. DIJKSTRA, *A note on two problems in connection with graphs*, Numer. Math., 1 (1959), pp. 269–271.

[5] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, San Francisco, 1979.

[6] S. L. HAKIMI, *Optimum locations of switching centers and the absolute centers and medians of a graph*, Operations Research, 12 (1964), pp. 450–459.

[7] G. Y. HANDLER AND P. B. MIRCHANDANI, *Location on Networks: Theory and Algorithms*, MIT Press, Cambridge, MA and London, 1979.

[8] E. L. LAWLER, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart and Winston, New York, 1976.

[9] F. ROMANI, *Shortest-path problem is not harder than matrix multiplication*, Information Processing Letters, 11 (1980).

[10] ———, private communication, 1981.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0009 $01.00/0

THE INTERVAL COUNT OF A GRAPH*

R. LEIBOWITZ†, S. F. ASSMANN‡ AND G. W. PECK‡

Abstract. The interval count of an interval graph G is the minimum number of different interval sizes needed to represent the vertices of G, where two vertices are adjacent if and only if their intervals intersect.

We show that if G is an interval graph and for some vertex x, $G-\{x\}$ has interval count one, then G has interval count two or less.

We also show how to construct examples of interval graphs where the interval count of G exceeds that of $G-\{x\}$ by at least two when the latter number is two or more.

1. Introduction. A graph G is an *interval graph* if each vertex can be assigned an interval of the real line in such a way that two vertices are adjacent if and only if their intervals intersect. See Fig. 1. Several characterizations of interval graphs have been given by various authors (Lekkerkerker and Boland [5], Gilmore and Hoffman [3], Fulkerson and Gross [2]), and they can be recognized in linear time (Booth and Leuker [1]). Such graphs have many applications.

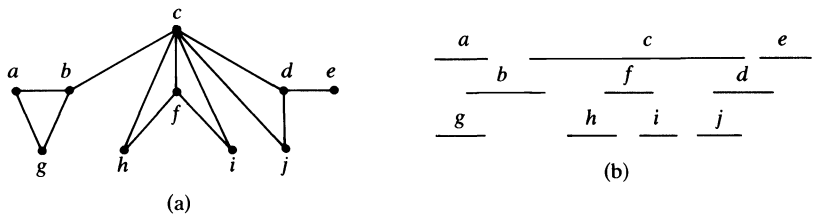


FIG. 1. *(a) An interval graph G. (b) A representation of G by intervals.*

R. L. Graham has suggested the general question of how many different sizes of intervals are necessary to represent a given interval graph G. The minimum number of sizes needed is called the *interval count* of G. In this paper we address a conjecture of Graham's which states that if the interval count of $G-\{x\}$ is k for some vertex x of G, then the interval count of G is at most $k+1$. We show that the conjecture is true for $k=1$ and false for $k \geqq 2$.

The next section contains some simple observations and techniques relating to interval representations of graphs. The third section contains a constructive proof of the conjecture for the case $k=1$, and the fourth section gives a method for producing counterexamples for the cases where $k \geqq 2$. We conclude with a brief discussion of the related open question of finding useful characterizations for those graphs with interval number k for fixed $k \geqq 2$.

2. Observations and techniques relating to interval representations. An interval representation of a graph is characterized by the orders of the left and right endpoints of the intervals. Suppose the left endpoints in order form the permutation π_L of the vertex labels and the right endpoints form the permutation π_R. If there is only one interval size, π_L and π_R must be the same. Conversely, if $\pi_L = \pi_R$, we can choose the intervals to all have the same size.

* Received by the editors June 13, 1979, and in final revised June 15, 1981.

† Department of Mathematics, Wheaton College, Norton, Massachusetts 02766.

‡ Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

When $\pi_L \neq \pi_R$, any pair of vertices that are ordered differently in the two permutations must have one interval contained in the other, as $L_y < L_z < R_z < R_y$. We say then that y *dominates* z.

Roberts [7] has shown that the graphs with interval count one are those interval graphs which do not contain $K_{1,3}$ (also called a 3-*claw*) as an induced subgraph. It is clear that $K_{1,3}$ cannot be represented using only one interval size. Conversely, suppose G is an interval graph with interval count at least 2. Pick a representation of G with the fewest pairs of vertices out of order. Suppose vertex z is dominated by vertex y. We could extend the interval representing z to the left until it reaches beyond the interval representing y, unless we were stopped by encountering the interval of a vertex u adjacent to y but not to z. Similarly, there must be another vertex v adjacent to y but not to z that prohibits us from extending the interval for z to the right. Thus, y, z, u and v form a 3-claw with apex y.

Given a representation of G, a *segment* is an interval of the real line which does not contain the left or right endpoint of any interval representing a vertex of G. Unless we specifically state otherwise, the segments we will be discussing are those which are contained in at least one interval representing a vertex of G. Suppose we are given an interval representation of a graph which uses only one size of interval and that some segment S of the representation is specified. We can increase the length of that segment until it accounts for nearly the total length of each interval which contains it. We call this a *blow up* of segment S. We can then blow up the segments closest to S which are not contained in intervals containing S, and so on, until all intervals are once again the same size but with their size mainly contained in the blown up segments.

Similarly, we can blow up m specified segments within a given interval until each accounts for nearly $1/m$ of the length of the interval and then blow up segments in succession until all intervals are the same size again. See Fig. 2.

3. Proof of the conjecture for $k = 1$. Suppose G is an interval graph with interval count greater than one but $G - \{x\}$ has interval count one. This means that every induced 3-claw of G includes the vertex x. Choose a representation of G with fewest pairs of vertices out of order between π_L and π_R. There are three cases: (1) x appears

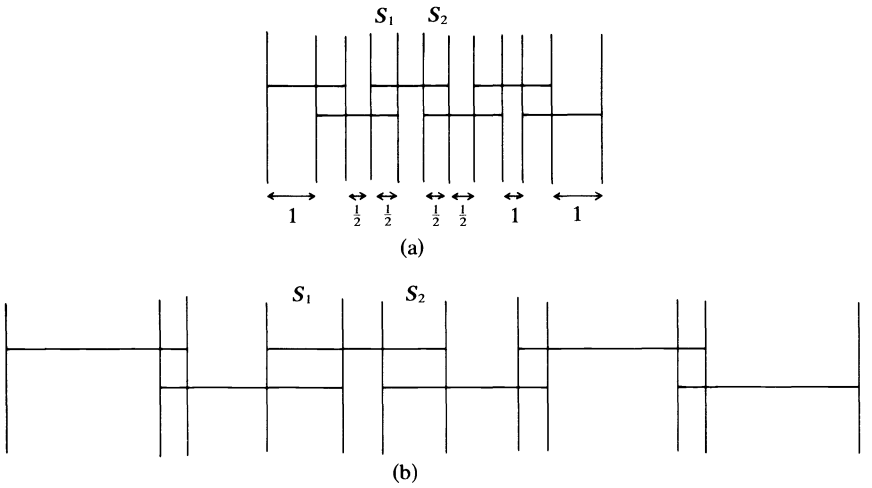


FIG. 2. (a) *A* 1*-sized representation before blowing up segments* S_1, S_2 *and successive segments. The segments to be blown up and the approximate amount of the total interval which each such segment is to account for are shown.* (b) *The* 1*-sized representation which results from the above operation.*

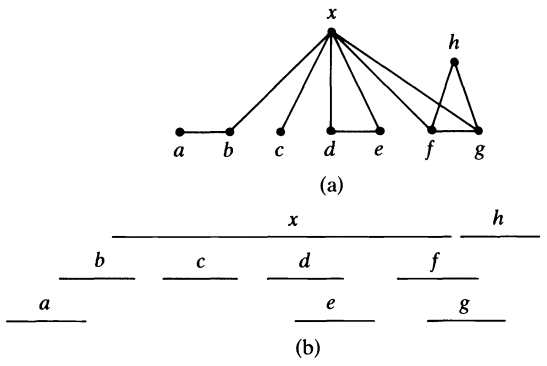


FIG. 3. *Case* 1. (a) *The graph G.* (b) *A representation of G using two interval sizes.*

in 3-claws only as an apex; (2) there is at least one vertex which dominates x; and (3) x is not dominated by any vertex but is a son in at least one 3-claw. (In this case x may also be an apex of some 3-claws.) In each case we will show how to construct a 2-sized representation for G.

Case 1. The only out of order pairs will be those where a vertex is dominated by x. Thus, we can make all intervals the same size except for x, which will have the second length. See Fig. 3.

Case 2. Let B be the set of vertices which are apices of 3-claws that have x as a right son, B' be those which have x as a left son, A be those vertices dominated by some vertex in B, A' be those dominated by some vertex in B' and C be those vertices which dominate x. Note that we may have $B = A = \emptyset$ and/or $B' = A' = \emptyset$.

Step 1. Remove the interval corresponding to x.

Step 2. Since there are no 3-claws in the remaining graph, we can move the vertices in A to the right until they extend beyond their respective fathers in B. Similarly, we can move the vertices of A' to the left.

Step 3. Adjust the lengths of the resulting intervals so that they are all the same size.

We claim that any vertex $y \notin \{A \cup B \cup x\}$ whose interval includes the right endpoint of some vertex in A or B includes the right endpoint of all vertices in A or B. (In particular, this holds for any $y \in C$, because such a y dominates x and x is adjacent to everything in B.)

To prove this claim, suppose y includes the right endpoint of some $a_i \in A$ but not $b_j \in B$. This means that in the original representation the interval for a_i lies to the right of that for b_j. But then x can't be adjacent to b_j without being adjacent to a_i, which is a contradiction. So if y intersects some a_i on the right, it intersects all the vertices in B. Suppose y includes the right endpoint of some $b_j \in B$. Then it includes the right endpoint of all the elements in A which b_j dominated (since these now extend to the right of b_j), and hence, it intersects all $b_k \in B$ and so all $a_i \in A$. The analogous fact is true for A' and B'.

Step 4. If $B \neq \emptyset$, blow up to size nearly $\frac{1}{2}$ the segment common to all the elements of A, B and all such y. (This segment exists by the above claim.)

If $B = \emptyset$, blow up to size nearly $\frac{1}{2}$ the segment common to exactly those vertices whose intervals included the left endpoint of x.

Do the analogous thing on the other side of the graph, according to whether $B' = \emptyset$ or not, in such a way that everything in C has size 1.

Step 5. Blow up segments in succession so that *all* intervals are of size 1.

Step 6. Change each interval of A to be of size $\frac{3}{5}$ by moving its right endpoint $\frac{2}{5}$ to the left. The intervals of A are now properly contained in the respective intervals of B. Likewise, move the left endpoints of each element of A' $\frac{2}{5}$ to the right.

Step 7. Insert an interval of size $\frac{3}{5}$ into the representation to correspond to x. It will fit to the right of the rightmost element of A and to the left of the leftmost element of A'.

See Fig. 4.

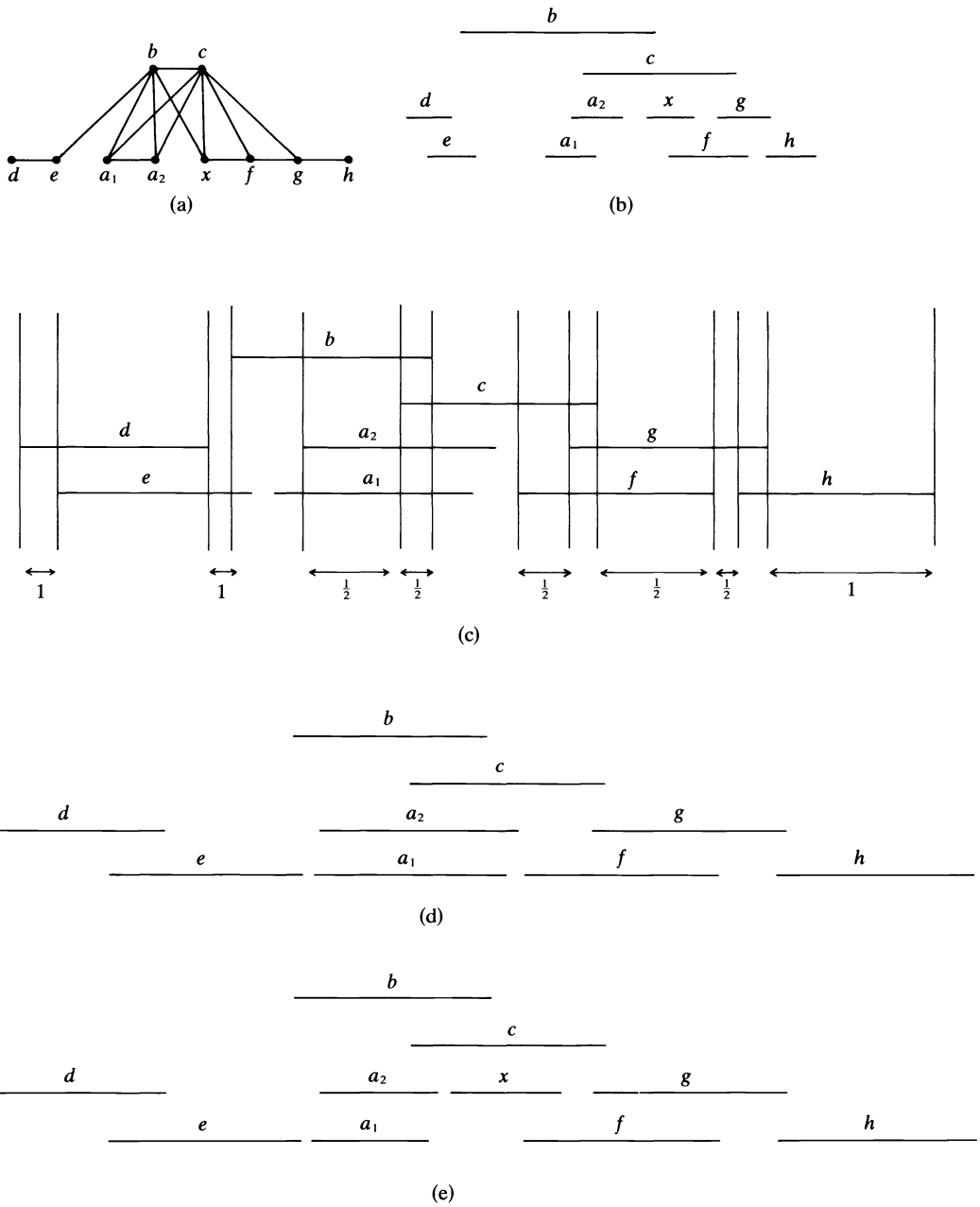


FIG. 4. *Case* 2. (a) *The graph G.* (b) *A representation of G.* (c) *After step* 3. (d) *After step* 5. (e) *The final,* 2*-sized representation.*

Case 3. Let B, A, B' and A' be defined as in Case 2. Note that at least one of B and B' (and, hence, A or A') is nonempty.

Step 1–*Step* 3. Same as in Case 2.

Step 4. Let C be the set of intervals which contained the left endpoint of x but did not correspond to x or to vertices in B. Let C' be defined similarly. If $C \neq \emptyset$, there is some segment S common to all elements of C and to no other intervals. If $C = \emptyset$, let S be the segment which contains no part of any interval and lies just to the right of the rightmost interval which lay entirely to the left of x in the original representation. Define C' and S' analogously.

Simultaneously blow up S and S' so that they are each nearly $\frac{1}{2}$ the total length from the leftmost left endpoint to the rightmost right endpoint of $G-\{x\}$. The elements of C and C' will now each have length $\frac{1}{2}$. Note every element of B is adjacent to every element of C, because elements of B also included the left endpoint of x. Similarly for B' and C'.

Step 5. Extend the elements of B to size $\frac{1}{2}$ by moving the right endpoint to the right. Now the elements of B dominate the elements of A. Perform a similar operation on the elements of B'.

Step 6. Insert an interval of size $\frac{1}{2}$ into the representation to correspond to x. This interval will have approximately half of its length in the segment S and half in the segment S' and will dominate anything which lays between these two segments, as we desired.

Note that the larger size interval can be arbitrarily big compared to the smaller intervals in this case, because x may dominate any number of independent vertices. See Fig. 5.

4. Counterexamples for $k \geqq 2$. We will give a counterexample for the case $k = 2$, which will easily generalize to higher values of k. Without loss of generality, all intervals are open intervals.

We need to find a graph G which has interval count at least four, while $G-\{x\}$ for some x has interval count 2.

One way to force a graph to have interval count at least four is as follows. Suppose vertex v_0 is the apex of a 3-claw. Then there is some vertex v_1 which it dominates. If v_1 is also the apex of a 3-claw, it dominates some vertex v_2. If v_2 is in turn the apex of a 3-claw, then it dominates some v_3, so we have a chain of four intervals, each strictly containing the next, so at least four interval sizes will be necessary to represent G.

However, we can force an interval count of at least four without having a nest of four intervals.

Suppose for the following discussion that G has interval count three. Let the interval sizes be λ, μ and β (for little, medium and big). If v dominates t independent vertices, then the size of v must be at least $t\lambda$. On the other hand, suppose there is a path of length $r+1$ from a vertex strictly on the left of v to one strictly on the right and that this path does not contain v. If each vertex on this path is dominated by some vertex in G, then the size of v can be no more than $r\mu$. One can easily arrange incompatible bounds on β by using these structures.

Consider the graph in Fig. 6. This can be represented with only two interval sizes; see Table 1. The vertices a_3 to a_7 are dominated by b_1 to b_5, respectively. If b_1 to b_5 are also each dominated by some vertex, they must be of size μ, and a_3 to a_7 would then be of size λ. Vertex a_2 is strictly to the left of b_3, and a_8 is strictly to the right, so because of the path $a_2, \cdots, a_8$, we must have $\mu \leqq 5\lambda$.

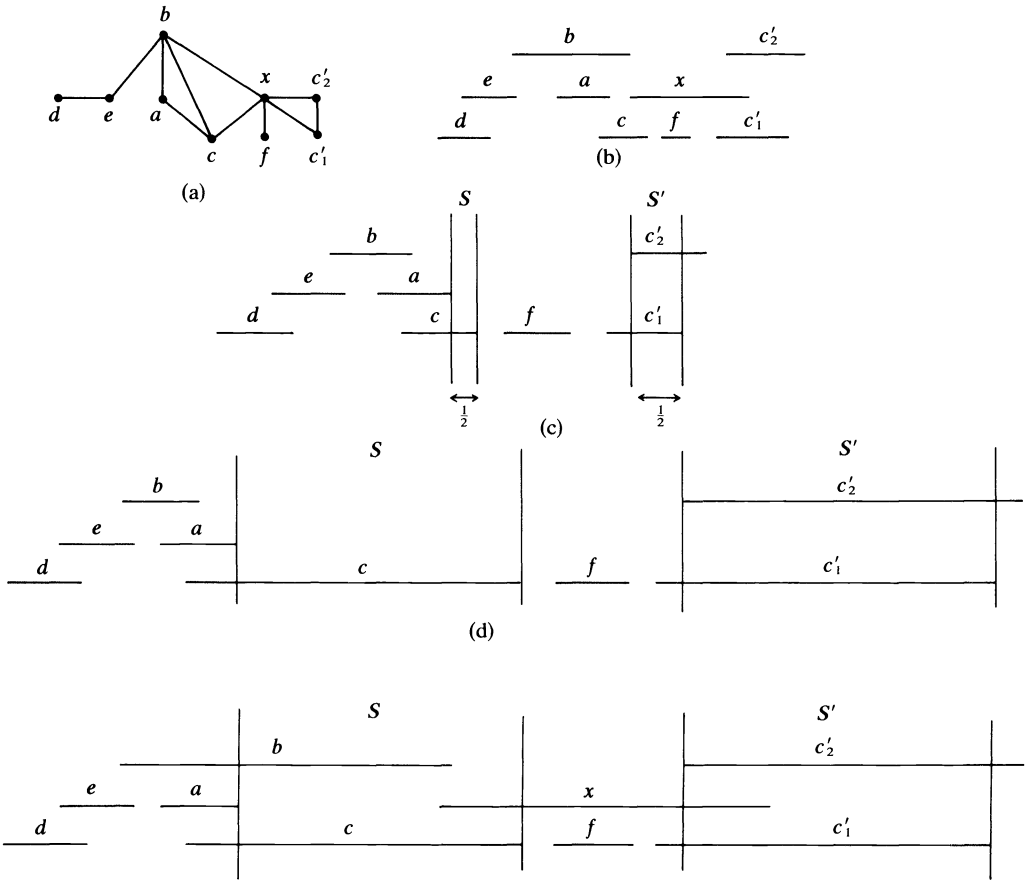


FIG. 5. *Case* 3. (a) *The graph G.* (b) *A representation of G.* (c) *After step* 3. (d) *After step* 4. (e) *The final, 2-sized representation.*

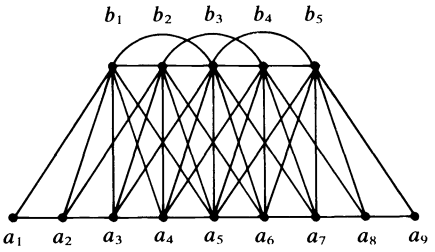


FIG. 6. *A component of the counterexample.*

Now consider the similar graph in Fig. 7. This can also be represented with only two interval sizes; see Table 2. Suppose we add a vertex x which is adjacent to c_1, c_2, y_0 to y_3 and u_1 *but not to* z. Vertex, u_1 dominates y_3 in the 3-claw u_1, y_1, y_3, y_5 and y_3 dominates z in the 3-claw y_3, x, z, u_3, so u_1 must have size β. One can easily check that x dominates y_0 and y_1, u_1 dominates y_2, u_2 dominates y_4 and u_3 dominates y_5, so the largest that y_0 through y_5 can be is size μ. Since c_2 is to the left of u_1, y_6 is to the right and y_0 to y_5 is a path between them, u_1 can be no larger than 6μ.

TABLE 1

A representation of the above graph using two interval sizes.

Vertex	Left endpoint	Right endpoint
a_1	0	72
a_2	36	108
b_1	54	162
a_3	72	144
b_2	90	198
a_4	108	180
b_3	136	244
a_5	144	216
b_4	162	270
a_6	180	252
b_5	198	306
a_7	216	288
a_8	252	324
a_9	288	360

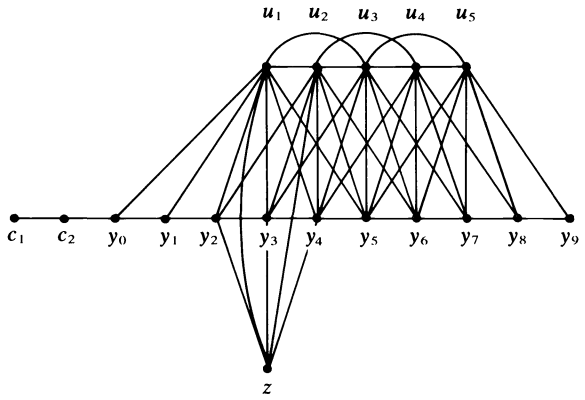


FIG. 7. *A component of the counterexample.*

Now consider the graph which consists of three disjoint copies of Fig. 6, one copy of Fig. 7 and a vertex x adjacent to all the points of each copy of Fig. 6, to c_1, c_2, y_0 through y_3 and u_1 and to 33 other points v_1 to v_{33}; see Fig. 8. Here the lines from x to some of the a_i's are not drawn in in order to avoid complicating the picture. This graph can be represented with four sizes of intervals; see Table 3. But suppose it can be represented with three sizes. Vertex x must have size at least 31λ, because it dominates v_2 through v_{32} and these are pairwise nonadjacent and so have nonoverlapping intervals. We have then that $\beta \geqq 31\lambda$. Because x, $b_{i,1}$, $b_{i,2}$, $b_{i,3}$ is a 3-claw for each $i = 1$ to 5, one set, say $b_{1,2}$ to $b_{5,2}$, must have each member dominated by x. Combining the above results, we get $\mu \leqq 5\lambda$ and $\beta \leqq 6\mu$, so $\beta \leqq 30\lambda$. Thus, assuming G can be represented with only three interval sizes leads to a contradiction.

TABLE 2

A representation of the graph in Fig. 7 using two interval sizes.

Vertex	Left endpoint	Right endpoint
c_1	0	72
c_2	3	75
y_0	72	144
y_1	75	147
u_1	143	251
y_2	144	216
z	171	243
y_3	178	250
u_2	192	300
y_4	216	288
u_3	249	357
y_5	250	322
u_4	255	363
y_6	288	360
u_5	306	414
y_7	324	396
y_8	360	432
y_9	396	468

TABLE 3

A representation of the graph in Fig. 8 using four sizes of intervals.

Vertices	Endpoints
$a_{1,1}$ to $a_{9,1}$ and $b_{1,1}$ to $b_{5,1}$	As in Table 1
$a_{1,2}$ to $a_{9,2}$ and $b_{1,2}$ to $b_{5,2}$	As in Table 1 plus 400
$a_{1,3}$ to $a_{9,3}$ and $b_{1,2}$ to $b_{5,3}$	As in Table 1 plus 800
v_1 to v_{33}	Left end of $v_i = 72i + 1128$
	Right end of $v_i = 72i + 1200$
c_1, c_2, y_0 to y_9, and u_1 to u_5	As in Table 2 plus 3600
z	Left end: 3784
	Right end: 3820
x	Left end: 1
	Right end: 3780

5. Given the simple characterization of graphs having interval count one, it is natural to seek such a characterization of graphs with interval count 2 (or k). Few results are known, but Leibowitz [6] has shown that graphs with interval count at most two include all trees which are interval graphs and all threshold graphs.

Golumbic [4] also mentions the open problem of finding good upper and lower bounds for the interval count of a graph.

Acknowledgment. One of the authors (Peck) would like to thank D. J. Kleitman for his many helpful comments.

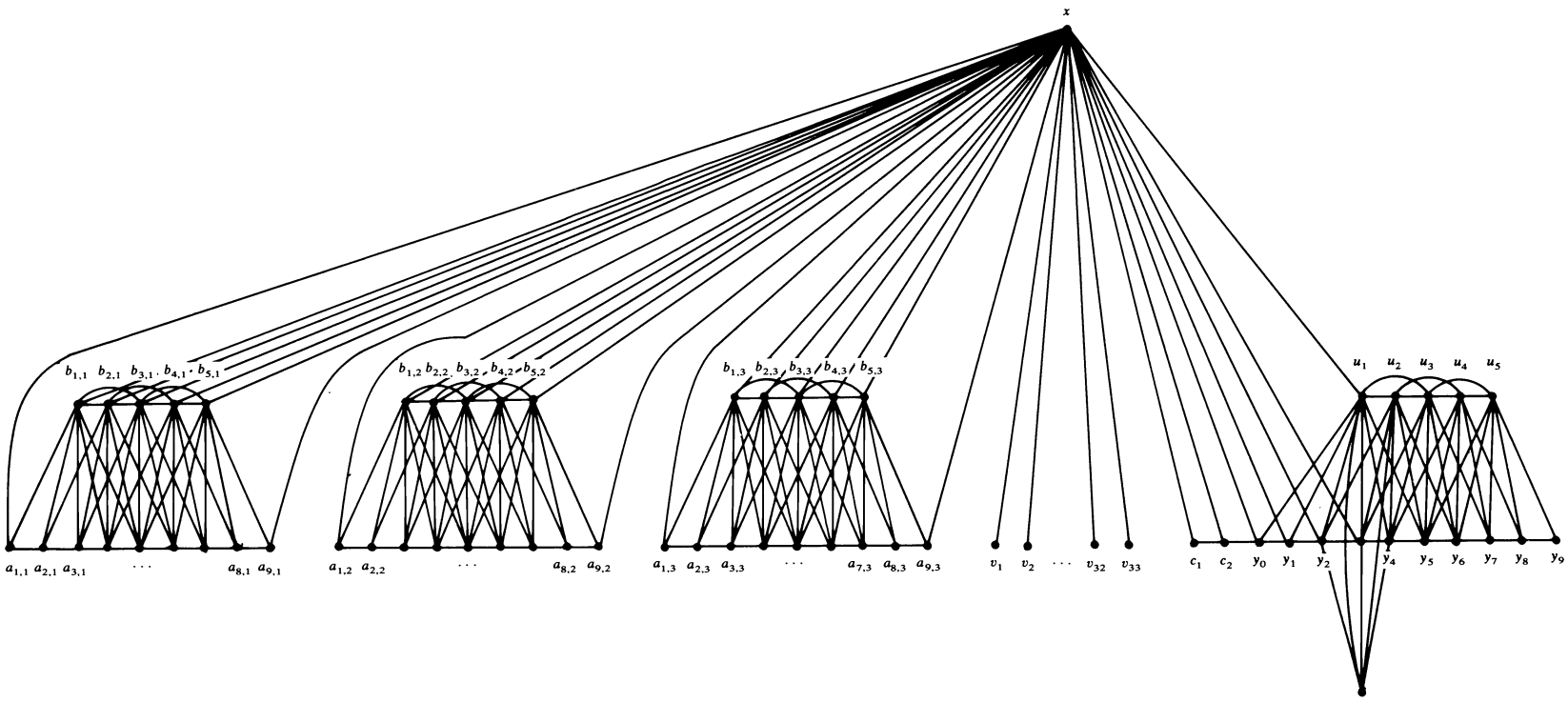


FIG. 8. *A counterexample for* $k = 2$.

REFERENCES

[1] KELLOGG S. BOOTH AND GEORGE S. LEUKER, *Testing for the consecutive ones property, interval graphs and graph planarity using PQ-tree algorithms*, J. Comput. Syst. Sci., 13 (1976), pp. 335–379.

[2] D. R. FULKERSON AND O. A. GROSS, *Incidence matrices and interval graphs*, Pacific J. Math., 15 (1965), pp. 835–855.

[3] P. C. GILMORE AND A. J. HOFFMAN, *A characterization of comparibility graphs and of interval graphs*, Canad. J. Math., 16 (1964), pp. 539–548.

[4] MARTIN CHARLES GOLUMBIC, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[5] C. G. LEKKERKERKER AND J. CH. BOLAND, *Representation of a finite graph by a set of intervals on the real line*, Fund. Math., 51 (1962), pp. 45–64.

[6] ROCHELLE LEIBOWITZ, *Interval counts and threshold graphs*, PhD. thesis, Rutgers University, New Brunswick, NJ.

[7] F. S. ROBERTS, *Indifference graphs*, in Proof Techniques in Graph Theory, F. Harary, ed., Academic Press, New York, 1969, pp. 139–146.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0010 $01.00/0

MAJORIZATION ON FINITE PARTIALLY ORDERED SETS*

KO-WEI LIH†

Abstract. The classical concept of majorization between two finite sequences of real numbers is extended to between real-valued functions defined on a finite partially ordered set. We establish characterizations of majorization. The FKG and Holley inequalities from statistical mechanics have their majorization interpretations. Their validity is closely tied to the distributiveness of the background lattice. An equivalence proof is hence provided. Finally, suitable restrictions on the rearrangement of function values enable us to generalize the classical theorem of Schur–Ostrowski which characterizes functions preserving majorization in terms of relative orders of their first partial derivatives.

1. Introduction. The classical concept of majorization between two finite sequences of real numbers is defined as follows. Let the sequences concerned be $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_n\}$ and $\beta = \{\beta_1, \beta_2, \cdots, \beta_n\}$. We rearrange them into nonincreasing order so that $\alpha_1^* \geqq \alpha_2^* \geqq \cdots \geqq \alpha_n^*$ and $\beta_1^* \geqq \beta_2^* \geqq \cdots \geqq \beta_n^*$, where α_i^* and β_i^* are, respectively, the ith largest numbers in α and β. Then α is said to *majorize* β if

$$\alpha_1^* + \alpha_2^* + \cdots + \alpha_k^* \geqq \beta_1^* + \beta_2^* + \cdots + \beta_k^*$$

for $k = 1, 2, \cdots, n$ and the equality holds when $k = n$. This definition is due to Muirhead [12] and is instrumental in the study of symmetric means by Hardy, Littlewood, and Pólya [5]. Subsequently, majorization was applied to a wide variety of problems such as deterministic and probabilistic inequalities, incidence matrices, order statistics, and optimal codings, just to name a few. The recent book by Marshall and Olkin [11] provides an excellent account of the many facets of majorization.

One slightly unpleasant feature of classical majorization is that it is not a partial ordering unless all rearrangements of a sequence are identified or only monotone sequences are considered. The rearrangement procedure impedes a direct and natural generalization to sequences defined on a finite partially ordered set. With the rearrangement step removed, it is convenient to consider α_i and β_i as weights attached to the element p_i of a finite partially ordered set P. And α is again said to *majorize* β if

$$\sum \{\alpha_i | p_i \in U\} \geqq \sum \{\beta_i | p_i \in U\}$$

for any subset U of P which is closed above and the equality holds in case $U = P$. When P is a totally ordered set $p_1 > p_2 > \cdots > p_n$, this new concept of majorization coincides with the classical one imposed upon nonincreasing sequences. Inspired by his research into Huffman trees [3], [8], Hwang [7] defines the above notion of majorization and attempts a generalization of the following fundamental theorem of Schur [15] and Ostrowski [13].

THEOREM. *Let $f(x_1, x_2, \cdots, x_n)$ be a real-valued function possessing continuous first partial derivatives. Then $f(\alpha_1, \alpha_2, \cdots, \alpha_n) \geqq f(\beta_1, \beta_2, \cdots, \beta_n)$ for all sequences α and β such that α majorizes β in the classical sense if and only if f is symmetric and $(x_i - x_j)((\partial f/\partial x_i) - (\partial f/\partial x_j)) \geqq 0$ holds in the domain of f.*

However, Hwang is only partly successful due to the lack of a suitable modification of the rearrangement procedure. Nevertheless, along the way, Hwang proves a beautiful theorem [7, Thm. 3.1] which characterizes majorization in terms of a finite sequence

* Received by the editors January 30, 1981. This work was supported in part by the National Science Council of the Republic of China.

† Institute of Mathematics, Academia Sinica, Taipei, Taiwan, Republic of China.

of "flows". His proof contains a grain of surprise because the main tool used is a result of Shapley [17] concerning the existence of cores of characteristic function games. Although interesting in its own right, this proof more or less obscures the underlying dynamics. Therefore, Hwang [9] subsequently provides a direct proof of his generalization of the Schur–Ostrowski theorem.

Our efforts to understand Hwang's flow characterization have guided our attention to an apparently remote topic—the inequality of Fortuin–Kasteleyn–Ginibre [4] from statistical mechanics. The FKG inequality states as follows. Let L be a finite distributive lattice and let μ be a probability measure on L satisfying

$$\mu(x)\mu(y) \leqq \mu(x \vee y)\mu(x \wedge y)$$

for all $x, y \in L$, where $\mu(x) = \mu(\{x\})$. Then

$$\sum_x f(x)g(x)\mu(x) \geqq \Big(\sum_x f(x)\mu(x)\Big)\Big(\sum_x g(x)\mu(x)\Big)$$

for real-valued functions f and g monotone in the sense of the order of L. The FKG inequality is known to have many combinatorial applications [16], [18] and to be a consequence of the following inequality of Holley [6]. Let μ_1 and μ_2 be probability measures on a finite distributive lattice L such that

$$\mu_1(x)\mu_2(y) \leqq \mu_1(x \vee y)\mu_2(x \wedge y)$$

for all $x, y \in L$. Then

$$\sum_x f(x)\mu_1(x) \geqq \sum_x f(x)\mu_2(x)$$

for any increasing function f. The Holley theorem so phrased is equivalent to another formulation within a network-flow setting as demonstrated in Preston [14]. Seymour and Welsh [16] also note that the equivalence proof is based on an order relation between μ_1 and μ_2 which is precisely the majorization relation on a partially ordered set introduced by Hwang. This strongly suggests the feasibility of a network-flow foundation for Hwang's flow characterization. Such are the circumstances which prompt us to embark upon a unification under the banner of majorization.

Now a few words about the organization of this paper are in order. In § 1, we first define basic terms and introduce the notions of majorization and weak majorization. Various characterizations of these notions are supplied afterwards. Theorem 2 is a modification of the original equivalent formulation of the Holley inequality. Theorem 3, generalizing Hwang's flow characterization, can be regarded as a "sequential" version of Theorem 2. Theorem 4 and its second corollary will bear out the majorization base of Holley and FKG inequalities. The phenomena that local order relations on the four-element sublattices $\{x, y, x \vee y, x \wedge y\}$ determine global majorization relations are closely tied to the distributiveness of the lattice. After introducing an auxiliary concept of majorization in ratio, the aforesaid phenomena are codified into the names of Holley and FKG lattices. Thereby their equivalence to distributiveness is established. An analogous result of Kemperman [10, Thm. 7] could be compared with our Theorem 6. In § 3, a well-behaved method of rearrangements enables us to obtain a bona fide generalization of the classical Schur–Ostrowski theorem. The proof also corroborates the usefulness of the flow characterization of majorization.

2. Characterization theorems. Let P be a finite partially ordered set. A subset U of P is called an (*order-*) *filter* if, for $x \in U$ and $y \in P$, $x \leqq y$ implies $y \in U$.

$\langle S\rangle = \{x \mid (\exists y \in S)(x \geqq y)\}$ is the filter generated by a subset S of P. We simply write $\langle x\rangle$ when S is the singleton set $\{x\}$. The set of all real-valued functions defined on P is denoted by $\mathscr{F}$. Lower Greek letters represent elements of $\mathscr{F}$. For α in $\mathscr{F}$ and $S \subseteq P$, we define $\alpha(S) = \sum\{\alpha(x) \mid x \in S\}$, when S is nonempty, and $\alpha(\varnothing) = 0$. A σ in $\mathscr{F}$ is called *increasing* if $\sigma(x) \geqq \sigma(y)$ whenever $x \geqq y$ in P.

A natural partial ordering of $\mathscr{F}$, named *weak majorization*, can be defined as follows.

α *weakly majorizes* β if and only if $\alpha(U) \geqq \beta(U)$ for any filter U.

The reflexivity and transitivity of weak majorization are immediate. Now suppose $\alpha(U) = \beta(U)$ for any filter U. Let x be an arbitrary element of P. Consider the filters $\langle x\rangle$ and $\langle x\rangle - \{x\}$. We see that $\alpha(x) = \alpha(\langle x\rangle) - \alpha(\langle x\rangle - \{x\}) = \beta(\langle x\rangle) - \beta(\langle x\rangle - \{x\}) = \beta(x)$. Thus weak majorization satisfies antisymmetry.

Now a stronger partial ordering of $\mathscr{F}$, called *majorization*, is defined as follows.

α *majorizes* β if and only if α weakly majorizes β and $\alpha(P) = \beta(P)$.

A totally ordered set T is said to be *consistent with* P if T and P possess exactly the same elements and $x \leqq y$ in P implies $x \leqq y$ in T. A characterization of majorization via totally ordered sets consistent with P is first exhibited in Hwang [9].

THEOREM 1. *α (weakly) majorizes β on P if and only if α (weakly) majorizes β on every totally ordered set T consistent with P.*

Proof. Sufficiency. Suppose $\alpha(U) < \beta(U)$ for some filter U with m elements. Since elements outside U are either incomparable or less than elements in U, there exists a totally ordered set T consistent with P whose m largest elements constitute the set U. It follows that α does not (weakly) majorizes β on T.

Necessity. Suppose $\sum_{i=1}^{m} \alpha(x_i) < \sum_{i=1}^{m} \beta(x_i)$ for the m largest elements of some totally ordered set T consistent with P. For $x \in P$, $x \geqq x_j$ in P implies $x \geqq x_j$ in T. So x must be some x_k, $1 \leqq k \leqq j$. This shows that $U = \{x_1, x_2, \cdots, x_m\}$ is a filter in P and $\alpha(U) < \beta(U)$. □

THEOREM 2. *Let α and β be nonnegative functions. α majorizes β if and only if there exists a nonnegative real-valued function γ defined on $P \times P$ such that*

$$\sum_{y \in P} \gamma(x, y) = \alpha(x), \quad \sum_{x \in P} \gamma(x, y) = \beta(y), \quad \gamma(x, y) = 0 \quad \textit{unless } x \geqq y.$$

A sufficient and necessary condition for α weakly majorizing β reads the same except that the first equality is replaced by $\sum_{y \in P} \gamma(x, y) \leqq \alpha(x)$.

Proof. We only treat the majorization case.

Sufficiency. Let U be an arbitrary filter. We also use U to denote the characteristic function of U. We have

$$\alpha(U) = \sum_x \sum_y U(x)\gamma(x, y) = \sum_y \sum_x U(x)\gamma(x, y) \geqq \sum_y \sum_x U(y)\gamma(x, y) = \beta(U).$$

Clearly $\alpha(P) = \beta(P)$. Thus α majorizes β.

Necessity. Construct a network-flow model as follows. Besides the source node s and the sink node t, we have nodes $(x, 1)$ and $(x, 2)$ for each $x \in P$. There is a directed edge from s to every $(x, 1)$ with capacity $\alpha(x)$. There is a directed edge from every $(x, 2)$ to t with capacity $\beta(x)$. There is a directed edge from $(x, 1)$ to $(y, 2)$ with unlimited capacity if and only if $x \geqq y$ in P. Suppose (S, T) is a cut with finite capacity. Let $T_i = \{x \in P \mid (x, i) \in T\}$ for $i = 1, 2$. The finiteness of the capacity forces $\langle T_2\rangle \subseteq T_1$. It follows that $\beta(T_2) \leqq \beta(\langle T_2\rangle) \leqq \alpha(\langle T_2\rangle) \leqq \alpha(T_1)$ and the cut with $T = \{t\}$ has the minimum

capacity. By the well-known max-flow-min-cut theorem, the maximum flow value of this network is equal to $\beta(P)$. Now define $\gamma(x, y)$ to be the value of a maximum flow on the directed edge from x to y if $x \geqq y$ in P. $\gamma(x, y)$ is defined to be 0 for all other pairs of elements. □

An α is said to be transformed into α' by a *reduction* if, for a certain pair $x \geqq y$ in P, we have

$$\alpha'(x) = \alpha(x) - \varepsilon, \quad \alpha'(y) = \alpha(y) + \delta, \quad \alpha'(z) = \alpha(z) \quad \text{for } z \neq x \text{ or } y,$$

where ε and δ are real numbers such that $\varepsilon \geqq 0$ and δ is equal to either ε or 0. If $\delta = \varepsilon$, the reduction is said to be *conservative*. α is *reducible* to β if α can be transformed into β by a finite sequence of reductions.

THEOREM 3. *α weakly majorizes β if and only if α is reducible to β. Moreover, α majorizes β if and only if α is reducible to β by a sequence of conservative reductions. When α and β are both nonnegative, a reduction sequence can be constructed through nonnegative functions.*

Proof. Sufficiency is immediate from the transitivity of (weak) majorization since α (weakly) majorizes α' after a single reduction.

Necessity. Let us first deal with majorization. Suppose α and β are nonnegative. Using Theorem 2, we can sequentially arrange conservative reductions between pairs $x > y$ with amounts $\varepsilon = \gamma(x, y)$. For arbitrary α and β, we may convert them into nonnegative functions by adding a sufficiently large constant to each $\alpha(x)$ and $\beta(x)$. The reduction amounts thereby obtained also work for the original α and β. To deal with weak majorization, we extend P to P' by adjoining a new element x_0 which is to be less than all elements of P. Extend α and β to P' by defining $\alpha(x_0) = \beta(P) - \alpha(P)$ and $\beta(x_0) = 0$. Thus $\alpha(P') = \beta(P')$. It is obvious that α majorizes β on P'. Hence α is reducible to β on P' by conservative reductions. If a reduction step involves $x > x_0$ and the amount ε, then it can be regarded as a reduction on P from x to x with ε subtracted and $\delta = 0$ added. Modified in this manner, a finite sequence of reductions from α to β on P is obtained. □

When P is a totally ordered set, we label elements of P in the standard order $p_1 > p_2 > \cdots > p_n$. Thus, an α can be identified with the row vector $(\alpha_1, \alpha_2, \cdots \alpha_n)$ such that $\alpha_i = \alpha(p_i)$. The following corollary becomes apparent when we interpret $m_{ij}\alpha_i$, $i < j$, as the reduction amount from p_i to p_j.

COROLLARY 3.1. *Let P be a totally ordered set. α majorizes β if and only if there exists an $n \times n$ matrix $M = (m_{ij})$ such that $m_{ij}\alpha_i \geqq 0$, $m_{ij} = 0$ for $i > j$, $\sum_{j=1}^{n} m_{ij} = 1$ and $\beta = \alpha M$.*

THEOREM 4. *α weakly majorizes β if and only if $\sigma\alpha$ weakly majorizes $\sigma\beta$ for any nonnegative increasing function σ. Here $\sigma\alpha$ is the product function such that $\sigma\alpha(x) = \sigma(x)\alpha(x)$ for all x.*

Proof. Sufficiency is trivial.

Necessity. Let α be transformed into α' by a reduction involving $x \geqq y$. Let U be an arbitrary filter.

$$\begin{aligned}\sigma\alpha(U) &= U(x)\sigma(x)(\alpha'(x) + \varepsilon) + U(y)\sigma(y)(\alpha'(y) - \delta) \\ &\quad + \sum \{\sigma(z)\alpha'(z) | z \in U - \{x, y\}\} \\ &= \varepsilon U(x)\sigma(x) - \delta U(y)\sigma(y) + \sigma\alpha'(U).\end{aligned}$$

(i) If $U(x) = U(y) = 0$, then $\sigma\alpha(U) = \sigma\alpha'(U)$.

(ii) If $U(x) = 1$ and $U(y) = 0$, then $\sigma\alpha(U) = \varepsilon\sigma(x) + \sigma\alpha'(U) \geqq \sigma\alpha'(U)$.

(iii) If $U(x) = U(y) = 1$, then $\sigma\alpha(U) = \varepsilon\sigma(x) - \delta\sigma(y) + \sigma\alpha'(U) \geqq \sigma\alpha'(U)$ since $\sigma(x) \geqq \sigma(y)$ and δ is either 0 or $\varepsilon \geqq 0$.

This shows that $\sigma\alpha$ weakly majorizes $\sigma\alpha'$. By Theorem 3, $\sigma\alpha$ weakly majorizes $\sigma\beta$. □

We remark that (i) the inequality $\sum_x \sigma(x)\alpha(x) \geqq \sum_x \sigma(x)\beta(x)$ is already sufficient and necessary for α weakly majorizing β; (ii) if α majorizes β, then, by adding a large enough constant to make σ nonnegative, we have $\sum_x \sigma(x)\alpha(x) \geqq \sum_x \sigma(x)\beta(x)$ for any increasing function σ.

COROLLARY 4.1. *Let $\mathcal{F}_1$ be the set of all nonnegative increasing functions and $\mathcal{F}_2$ be the set of all nonnegative σ such that, for any α and β, if α weakly majorizes β, then $\sigma\alpha$ weakly majorizes $\sigma\beta$. Then $\mathcal{F}_1 = \mathcal{F}_2$.*

Proof. Theorem 4 shows that $\mathcal{F}_1 \subseteq \mathcal{F}_2$. Conversely, let $\sigma \in \mathcal{F}_2$. For $x \geqq y$, let α and β be the characteristic functions of $\{x\}$ and $\{y\}$, respectively. It is clear that α weakly majorizes β. So $\sigma(x) = \sigma\alpha(P) \geqq \sigma\beta(P) = \sigma(y)$. □

COROLLARY 4.2. *Suppose $\sigma\alpha(P) \neq 0$. σ^α majorizes α if and only if $\tau\sigma^\alpha$ weakly majorizes $\tau\alpha$ for any nonnegative increasing function τ. Here σ^α is the average of σ with respect to α defined by $\sigma^\alpha(x) = (\alpha(P)/\sigma\alpha(P))\sigma(x)\alpha(x)$ for all x.*

This is immediate since $\sigma^\alpha(P) = \alpha(P)$. If $\sigma\alpha(P) \neq 0$ and σ^α majorizes α, we actually have $\tau\sigma^\alpha(P) \geqq \tau\alpha(P)$ for any increasing function τ. This can be verified by adding a sufficiently large constant to each $\tau(x)$ and applying Theorem 4.

3. Holley and FKG properties. We say that α *majorizes* β *in ratio* if $\beta(P)\alpha(U) \geqq \alpha(P)\beta(U)$ for any filter U. For nonnegative functions, α majorizing β in ratio together with $\alpha(P) \geqq \beta(P)$ will give rise to a partial ordering stronger than weak majorization. Other elementary facts concerning majorization in ratio are collected in the following lemma. The straightforward proof is omitted.

LEMMA 5. *Let α, β and γ be positive functions. Then the following statements are equivalent.*

(i) *α^γ majorizes β^γ.*

(ii) *$\gamma\alpha$ majorizes $\gamma\beta$ in ratio.*

(iii) *γ^α majorizes γ^β in ratio.*

Now P is assumed to be a lattice with the join $\vee$ and the meet $\wedge$ operations. We have the following notions for P.

P is said to be *Holley* if, when α and β are nonnegative, the inequality

$$\alpha(x)\beta(y) \leqq \alpha(x \vee y)\beta(x \wedge y)$$

for all x and y will imply that α majorizes β in ratio.

P is said to be FKG if, when α is nonnegative, the inequality

$$\alpha(x)\alpha(y) \leqq \alpha(x \vee y)\alpha(x \wedge y)$$

for all x and y will imply that σ^α majorizes α for any nonnegative increasing σ such that $\sigma\alpha(P) > 0$.

Note that, when P is FKG, the conditions on α imply

$$\left(\sum_x \alpha(x)\right)\left(\sum_x \sigma(x)\tau(x)\alpha(x)\right) \geqq \left(\sum_x \sigma(x)\alpha(x)\right)\left(\sum_x \tau(x)\alpha(x)\right)$$

for any increasing functions σ and τ.

THEOREM 6. *The following statements are equivalent for a finite lattice P.*

(i) *P is Holley.*

(ii) *P is* FKG.

(iii) *P is distributive.*

Proof. From (i) *to* (ii). Let α and σ be nonnegative, $\sigma\alpha(P) > 0$, and σ increasing.

Clearly σ^α is nonnegative. Assume $\alpha(x)\alpha(y) \leqq \alpha(x \vee y)\alpha(x \wedge y)$ for all x and y.

$$\begin{aligned}\sigma^\alpha(x)\alpha(y) &= (\alpha(P)/\sigma\alpha(P))\sigma(x)\alpha(x)\alpha(y)\\ &\leqq (\alpha(P)/\sigma\alpha(P))\sigma(x)\alpha(x \vee y)\alpha(x \wedge y)\\ &\leqq (\alpha(P)/\sigma\alpha(P))\sigma(x \vee y)\alpha(x \vee y)\alpha(x \wedge y)\\ &= \sigma^\alpha(x \vee y)\alpha(x \wedge y)\end{aligned}$$

Since P is Holley, σ^α majorizes α in ratio. Hence (ii) holds.

From (ii) *to* (iii). Let P_0 be a sublattice of the FKG lattice P. Assume that α and σ are nonnegative functions on P_0 such that σ is increasing on P_0 and $\sigma\alpha(P_0) > 0$. Now extend α and σ to α' and σ' which are defined on P. For $x \in P - P_0$, define $\alpha'(x) = 0$ and $\sigma'(x) = \max\{\sigma(y) | y \leqq x \text{ and } y \in P_0\}$ under the convention $\max \varnothing = 0$. It follows that $\alpha'(x)\alpha'(y) \leqq \alpha'(x \vee y)\alpha'(x \wedge y)$ for any $x, y \in P$ and σ' is increasing on P. Suppose U_0 is a filter in P_0. We consider the filter $U = \langle U_0 \rangle$ in P. Since P is FKG, we have $(\alpha'(P)/\sigma'\alpha'(P))\sigma'\alpha'(U) \geqq \alpha'(U)$. However, $\alpha'(P) = \alpha(P_0)$, $\alpha'(U) = \alpha(U_0)$, $\sigma'\alpha'(P) = \sigma\alpha(P_0)$, and $\sigma'\alpha'(U) = \sigma\alpha(U_0)$. Thus σ^α majorizes α on P_0, i.e., P_0 is also FKG. On the other hand, a lattice is distributive if and only if it has neither M_5 nor N_5 as a sublattice, where $M_5 = \{x_1 > x_2, x_3, x_4 > x_0\}$ and $N_5 = \{x_1 > x_2 > x_3 > x_0$ and $x_1 > x_4 > x_0\}$. To prove our implication, it suffices to show that neither M_5 nor N_5 is FKG. Now let α be the constant function $\frac{1}{5}$ on M_5. Let σ be the characteristic function of $\{x_1, x_2, x_3\}$. It follows that $\sigma^\alpha = \sigma/3$. On the filter $U = \{x_1, x_4\}$ we have $\sigma^\alpha(U) = \frac{1}{3} < \frac{2}{5} = \alpha(U)$. So σ^α does not majorize α. N_5 can be dealt with similarly. This example is first observed by Kemperman [10, p. 327] in a slightly different context.

From (iii) *to* (i). This is already established in Ahlswede and Daykin [2, p. 288] by means of their remarkable four-weight inequality in [1]. □

4. Schur–Ostrowski theorem. In order to generalize the classical Schur–Ostrowski theorem to majorization on partially ordered sets, we have to introduce a suitable notion for the rearrangement of values of a function. An α' is said to be obtained from α by a *transposition* if there exist x and y such that $\alpha'(x) = \alpha(y)$, $\alpha'(y) = \alpha(x)$ and $\alpha'(z) = \alpha(z)$ for $z \neq x$ and y. α' is a *rearrangement* of α if it can be obtained from α by a finite sequence of transpositions. α' is a *P-rearrangement* of α if each of the transpositions in the rearrangement sequence involves two comparable elements of P. We first note the following existence lemma.

LEMMA 7. *For any α, there exists an increasing P-rearrangement α^* of α such that α^* majorizes α.*

Proof. This can be established by induction on the cardinality of P. Let $\alpha(y) = \max\{\alpha(z) | z \in P\}$ and x be a maximal element in $\langle y \rangle$. By transposing values at x and y, we get a P-rearrangement α' of α. It is straightforward to verify that α' majorizes α. Now let α'' be the restriction of α' to $P' = P - \{x\}$. By induction, we have some increasing P'-rearrangement $(\alpha'')^*$ of α'' which majorizes α'' on P'. Restoring the omitted value to $(\alpha'')^*$, we get an increasing P-rearrangement α^* which majorizes α. □

Throughout this section, we use α^* to denote an arbitrary increasing P-rearrangement of α. Now we fix a labeling of elements of P so that $P = \{p_1, p_2, \cdots, p_n\}$. As noted before, when P is a totally ordered set, the labeling satisfies $p_1 > p_2 > \cdots > p_n$. An α is identified with the point $(\alpha_1, \alpha_2, \cdots, \alpha_n)$, with $\alpha_i = \alpha(p_i)$, in the n-dimensional real space. A real-valued function $f(x_1, x_2, \cdots, x_n)$ is said to be *P-symmetric* if, for any comparable elements p_i and p_j of P, the function value will be the same when x_i and x_j are interchanged. A short notation for the substitution $f(\alpha_1, \alpha_2, \cdots, \alpha_n)$ is $f(\alpha)$.

To state our generalization of the Schur–Ostrowski theorem, we need the set D_{ij} for any comparable elements p_i and p_j. $D_{ij} = \{\alpha \mid$there exists some α^* in which α_i and α_j are rearranged to comparable elements of $P\}$.

THEOREM 8. *Let $f(x_1, x_2, \cdots, x_n)$ be a function possessing continuous first partial derivatives over an open domain. Then $f(\alpha) \geqq f(\beta)$ for all α and β such that some α^* majorizes some β^*, if and only if f is P-symmetric and, for any comparable elements p_i and p_j,*

$$(x_i - x_j)\left(\frac{\partial f}{\partial x_i} - \frac{\partial f}{\partial x_j}\right) \geqq 0$$

on $D_{ij} \cap$ domain (f).

Proof. Necessity. Suppose p_i and p_j are comparable. Let $\beta_i = \alpha_j, \beta_j = \alpha_i$, and $\beta_k = \alpha_k$ for $k \neq i$ and j. Then any α^* can be regarded as a β^* by adding at most one transposition, and vice versa. So f is P-symmetric.

Now choose and fix an arbitrary $\alpha \in D_{ij} \cap$ domain (f). To simplify notation, we may let $i = 1$ and $j = 2$. For $\varepsilon > 0$, define β^ε as follows.

$$\beta_1^\varepsilon = (1-\varepsilon)\alpha_1 + \varepsilon\alpha_2, \quad \beta_2^\varepsilon = \varepsilon\alpha_1 + (1-\varepsilon)\alpha_2, \quad \beta_k^\varepsilon = \alpha_k \quad \text{for } k \neq 1 \text{ or } 2.$$

Now assume $\alpha_1 > \alpha_2$. Let α^* be the increasing P-rearrangement of α such that α_1 and α_2 are rearranged to comparable elements $p_s > p_t$. We may suppose that there is no p_k satisfying $p_s > p_k > p_t$ and either $\alpha_k^* = \alpha_1$ or $\alpha_k^* = \alpha_2$. Now, in α^*, we replace the α_1 at p_s by β_1^ε and the α_2 at p_t by β_2^ε. For all sufficiently small ε, the result of such replacements produces some $(\beta^\varepsilon)^*$ and shows $\beta^\varepsilon \in D_{12}$. Clearly α^* majorizes $(\beta^\varepsilon)^*$. Thus $f(\alpha) - f(\beta) \geqq 0$ for all sufficiently small ε. Using the mean value theorem, we see that

$$\begin{aligned}\left(\frac{1}{\varepsilon}\right)(f(\alpha) - f(\beta)) &= (\alpha_1 - \alpha_2)\left(\frac{\partial f}{\partial x_1}((1-\lambda\varepsilon)\alpha_1 + \lambda\varepsilon\alpha_2, \lambda\varepsilon\alpha_1 + (1-\lambda\varepsilon)\alpha_2, \alpha_3, \cdots, \alpha_n)\right. \\ &\qquad \left. - \frac{\partial f}{\partial x_2}((1-\lambda\varepsilon)\alpha_1 + \lambda\varepsilon\alpha_2, \lambda\varepsilon\alpha_1 + (1-\lambda\varepsilon)\alpha_2, \alpha_3, \cdots, \alpha_n)\right)\end{aligned}$$

for some λ, $0 < \lambda < 1$. As ε approaches 0, this shows that $(\alpha_1 - \alpha_2) \cdot (\partial f(\alpha)/\partial x_1 - \partial f(\alpha)/\partial x_2) \geqq 0$. The case for $\alpha_1 < \alpha_2$ can be handled similarly.

Sufficiency. We want to show that $f(\alpha) \geqq f(\beta)$ for all α and β such that some α^* majorizes some β^*. Since f is P-symmetric, $f(\alpha) = f(\alpha^*)$ and $f(\beta) = f(\beta^*)$. Therefore we may suppose that α and β are increasing and α majorizes β.

Assume, on the contrary, that, for a certain increasing α, there exists an increasing β such that α majorizes β and $f(\alpha) < f(\beta)$. Consider the set $C = \{\beta \mid \alpha$ majorizes β and β is increasing.$\}$. For any totally ordered set T consistent with P, we define $C_T = \{\beta \mid \alpha$ majorizes β on T and β is increasing on $P.\}$. It follows from Theorem 1 that $C = \cap C_T$, where the intersection is taken over those finitely many totally ordered sets consistent with P. The compactness of C in the n-dimensional real space follows from the compactness of each and every C_T.

We claim that C_T is bounded. Every β in C_T can be obtained from α by a finite sequence of conservative reductions. Thus, if one coordinate corresponding to p_i grows to infinity, then a certain coordinate corresponding to some $p_j > p_i$ will diminish to negative infinity. This produces nonincreasing functions. Similarly no coordinate diminishes to negative infinity.

Next claim that C_T is closed. Let $\beta^1, \beta^2, \cdots$ be a sequence in C_T which converges to β^0. We have to show $\beta^0 \in C_T$. Write $\beta^k = \alpha M_k$ where $M_k = (m_{ij}^k)$ are the matrices guaranteed by Corollary 3.1. Since each β^k is increasing, β^0 is increasing and the sequence m_{ij}^k, $k = 1, 2, \cdots$, is bounded for any i and j. By choosing appropriate subsequences, we may assume that $m_{ij}^0 = \lim_{k\to\infty} m_{ij}^k$ exists for all i and j. Then $M_0 = (m_{ij}^0)$ satisfies conditions listed in Corollary 3.1 and $\beta^0 = \alpha M_0$.

Consequently, f attains its maximum value on C at a certain β. β cannot be identical with α. Hence there exists a nontrivial sequence of conservative reductions $\alpha, \alpha_1, \alpha_2, \cdots, \alpha_t, \beta$ from α to β. Consider the last step from α_t to β. This step involves a pair of comparable elements, say $p_i > p_j$, and the amount $\varepsilon > 0$. Now define β^λ as follows.

$$\beta_i^\lambda = \beta_i + \lambda, \quad \beta_j^\lambda = \beta_j - \lambda, \quad \beta_k^\lambda = \beta_k \quad \text{for } k \neq i \text{ or } j,$$

where $0 \leqq \lambda \leqq \varepsilon$. We can easily modify the foregoing reduction sequence to become a reduction sequence from α to β^λ. Since β is increasing, we are able to show $\beta^\lambda \in D_{ij}$ for all λ less than a certain λ_0 by at most two transpositions. Furthermore, f still attains its maximum at β on the set $\{\beta^\lambda | 0 \leqq \lambda \leqq \lambda_0\}$. Let $g(\lambda) = f(\beta^\lambda)$. Thus $g'(y) = \partial f(\beta^\lambda)/\partial x_i - \partial f(\beta^\lambda)/\partial x_j$. We have two possibilities.

Case 1. On $D_{ij} \cap$ domain (f), we have

$$(x_i - x_j)\left(\frac{\partial f}{\partial x_i} - \frac{\partial f}{\partial x_j}\right) > 0$$

whenever $x_i \neq x_j$. Then

$$(\beta_i - \beta_j + 2\lambda)\left(\frac{\partial f(\beta^\lambda)}{\partial x_i} - \frac{\partial f(\beta^\lambda)}{\partial x_j}\right) > 0$$

implies $g'(\lambda) > 0$ which contradicts the maximality of $f(\beta)$. We hence obtain the desired inequality.

Case 2. *Otherwise.* A trick of Ostrowski will reduce this case to the previous one. In place of f, we consider the function $F(x_1, x_2, \cdots, x_n) = f(x_1, x_2, \cdots, x_n) + (\theta/2)(x_1^2 + x_2^2 + \cdots + x_n^2)$ for $\theta > 0$. We see that

$$(x_i - x_j)\left(\frac{\partial F}{\partial x_i} - \frac{\partial F}{\partial x_j}\right) = (x_i - x_j)\left(\frac{\partial f}{\partial x_i} - \frac{\partial f}{\partial x_j}\right) + \theta(x_i - x_j)^2 > 0$$

whenever $x_i \neq x_j$ on $D_{ij} \cap$ domain (f). By Case 1, $F(\alpha) \geqq F(\beta)$ for all α and β such that some α^* majorizes some β^*. Let θ approach 0; we have $f(\alpha) \geqq f(\beta)$. This completes the proof. □

In conclusion, we remark that Theorem 8 gives us the classical Schur–Ostrowski theorem when P is a totally ordered set. The rearrangement of sequences of real numbers can be achieved by transpositions each of which is trivially over two comparable elements of P. Furthermore, P-symmetry means symmetry and every D_{ij} is the set of all n term sequences. We further note that the proof of Theorem 8 can be modified slightly to obtain yet another generalization of the Schur–Ostrowski theorem.

THEOREM 9. *Let $f(x_1, x_2, \cdots, x_n)$ be a function possessing continuous first partial derivatives over an open domain. Then $f(\alpha) \geqq f(\beta)$ for all α and β such that some increasing rearrangement of α majorizes some increasing rearrangement of β, if and only if f is symmetric and, for any comparable elements p_i and p_j,*

$$(x_i - x_j)\left(\frac{\partial f}{\partial x_i} - \frac{\partial f}{\partial x_j}\right) \geqq 0$$

on $D'_{ij} \cap$ domain (f), where $D'_{ij} = \{\alpha \mid$ there exists some increasing rearrangement of α in which α_i and α_j are rearranged to comparable elements of $P\}$.

REFERENCES

[1] R. AHLSWEDE AND D. E. DAYKIN, *An inequality for the weights of two families of sets, their unions and intersections*, Z. Wahrsch. Verw. Gebiete, 43 (1978), pp. 183–185.

[2] ———, *Inequalities for a pair of maps $S \times S \to S$ with S a finite set*, Math. Z., 165 (1979), pp. 267–289.

[3] F. R. K. CHUNG AND F. K. HWANG, *On blocking probabilities on a class of linear graphs*, Bell System Tech. J., 57 (1978), pp. 2915–2925.

[4] C. M. FORTUIN, P. W. KASTELEYN AND J. GINIBRE, *Correlation inequalities on some partially ordered sets*, Comm. Math. Phys., 22 (1971), pp. 89–103.

[5] G. H. HARDY, J. E. LITTLEWOOD AND G. PÓLYA, *Inequalities*, 2nd ed., Cambridge Univ. Press, New York, 1959.

[6] R. HOLLEY, *Remarks on the FKG inequalities*, Comm. Math. Phys., 36 (1974), pp. 227–237.

[7] F. K. HWANG, *Majorization on a partially ordered set*, Proc. Amer. Math. Soc., 76 (1979), pp. 199–203.

[8] ———, *Generalized Huffman trees*, SIAM J. Appl. Math., 37 (1979), pp. 124–127.

[9] ———, *Generalized Schur functions*, Bull. Inst. Math. Acad. Sinica, 8 (1980), pp. 513–516.

[10] J. H. B. KEMPERMAN, *On the FKG-inequality for measures on a partially ordered space*, Indag. Math., 39 (1977), pp. 313–331.

[11] A. W. MARSHALL AND I. OLKIN, *Inequalities: Theories of Majorizations and Its Applications*, Academic Press, New York, 1979.

[12] R. F. MUIRHEAD, *Some methods applicable to identities and inequalities of symmetric algebraic functions of n letters*, Proc. Edinburgh Math. Soc., 21 (1903), pp. 144–157.

[13] P. A. OSTROWSKI, *Sur quelques applications des fonctions convexes et concaves au sens de I. Schur*, J. Math. Pures Appl., 31 (1952), pp. 253–292.

[14] C. J. PRESTON, *A generalization of the FKG inequalities*, Comm. Math. Phys., 36 (1974), pp. 233–241.

[15] I. SCHUR, *Über ein Klasse von Mittelbildungen mit Anwendungen auf die Determinantetheorie*, Sitzungsber. Berlin. Math. Ges., 22 (1923), pp. 9–20.

[16] P. D. SEYMOUR AND D. J. A. WELSH, *Combinatorial applications of an inequality from statistical mechanics*, Math. Proc. Cambridge Philos. Soc., 77 (1975), pp. 485–495.

[17] L. S. SHAPLEY, *Cores of convex gemes*, Internat. J. Game Theory, 1 (1971), pp. 11–26.

[18] L. A. SHEPP, *The FKG inequality and some monotonicity properties of partial orders*, this Journal, 1 (1980), pp. 295–299.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0011 $01.00/0

A PRIMAL APPROACH TO THE SIMPLE PLANT LOCATION PROBLEM*

GERARD CORNUEJOLS† AND JEAN-MICHEL THIZY‡

Abstract. The most successful algorithms for solving simple plant location problems are presently dual-based procedures. However, primal procedures have distinct practical advantages (e.g., in sensitivity analysis). We propose a primal subgradient algorithm to solve the well-known strong linear programming relaxation of the problem. Typically this algorithm converges very fast to a point whose objective value is close to the integer optimum and where most of the decision variables have been fixed either to 0 or to 1. To fix the values of the remaining variables we use a greedy-interchange algorithm. Thus we propose this approach as a heuristic. Computational experience shows that an optimal solution is discovered with high frequency.

1. The simple plant location problem. The simple plant location problem has received much attention, as it combines interesting theoretical features with a wide range of practical applications. Numerous algorithms have successively compounded on the various structural properties of its primal and dual linear programming relaxations. Although primal approaches have distinct advantages (e.g., in sensitivity analysis) they have so far been outperformed by dual approaches, e.g., Erlenkotter [5]. Geoffrion [6], Narula, Ogbu and Samuelsson [10] and Cornuejols, Fisher and Nemhauser [2] proposed a subgradient procedure on a dual formulation expressed as the optimization of a piecewise linear function. In this paper we show that the primal problem can be written in a similar form and solved by a subgradient algorithm. This primal approach presents the double advantage of directly yielding (near) optimal solutions which are amenable to 0–1 variable treatments and of being extremely easy to program. It has often been noted in the literature that subgradient methods yield good suboptimal solutions fairly rapidly, but are less efficient in a strict optimality search. We observed this behavior here also for the test problems that we tried. Accordingly, the approach proposed in this paper is to supplement the subgradient algorithm by a heuristic instead of continuing the search until optimality.

The simple plant location problem consists in selecting which facilities to keep open among a finite set J of potential sites in order to maximize the profit made in serving a finite set I of customers. A revenue c_{ij} can be made by satisfying the totality of the demand of customer i from location j if a facility is kept open there, and a cost f_j is incurred for maintaining a facility open at location j.

$$z^* = \max \sum_{i \in I} \sum_{j \in J} c_{ij} x_{ij} - \sum_{j \in J} f_j y_j, \tag{1}$$

$$\sum_{j \in J} x_{ij} = 1 \qquad \text{all } i \in I, \tag{2}$$

$$x_{ij} \leqq y_j \leqq 1 \qquad \text{all } i \in I, j \in J, \tag{3}$$

$$x_{ij}, \quad y_j \geqq 0 \qquad \text{all } i \in I, j \in J, \tag{4}$$

$$y_j \text{ integer} \qquad \text{all } j \in J. \tag{5}$$

* Received by the editors November 30, 1981. This research was supported by the National Science Foundation under grant ENG 7902506.

† Center for Operations Research and Econometrics, Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium.

‡ Departement Werktuigkunde, Industrieel Beleid, Katholieke Universiteit Leuven, 3030 Heverlee-Leuven, Belgium.

By relaxing the integrality conditions (5) we obtain the well-known *strong linear programming relaxation* (1)–(4). Several authors have observed that the strong linear programming relaxation often has integral optimal solutions [13], [2], [5]. We will show that the vector of variables y can be considered strategic in the sense that, when this vector is known, optimal values for the other variables (namely the variables x_{ij}) can be obtained analytically. Therefore, for any given feasible y, the optimal value of (1)–(4) takes an analytic form $z_L(y)$, and the strong linear programming relaxation reduces to

$$z^* = \max_{0 \leqq y \leqq 1} z_L(y).$$

PROPOSITION 1. *Let $J = \{1, 2, \cdots, n\}$ and, for each $i \in I$, let s_i be a permutation of J such that*

$$c_{is_i(1)} \geqq c_{is_i(2)} \geqq \cdots \geqq c_{is_i(n)}.$$

Given a vector $0 \leqq y \leqq 1$, an optimal solution to the linear program (1)–(4) *with variables x_{ij} is given by*

$$x_{is_i(j)} = \begin{cases} y_{s_i(j)} & \text{for } j < h_i, \\ 1 - \sum_{t<h_i} y_{s_i(t)} & \text{for } j = h_i, \\ 0 & \text{for } j > h_i, \end{cases} \tag{6}$$

where h_i is an index such that

$$\sum_{t<h_i} y_{s_i(t)} \leqq 1 \leqq \sum_{t \leqq h_i} y_{s_i(t)}. \tag{7}$$

Furthermore, the optimal value $z_L(y)$ is a piecewise linear concave function of y.

Proof. Given the vector $0 \leqq y \leqq 1$, the linear program (1)–(4) reduces to

$$\max \sum_{i \in I} \sum_{j \in J} c_{ij} x_{ij},$$

$$\sum_{j \in J} x_{ij} = 1 \quad \text{all } i \in I,$$

$$0 \leqq x_{ij} \leqq y_i \quad \text{all } i \in I, j \in J.$$

This program decomposes into $|I|$ knapsack problems with real bounded variables:

$$z_i(y) = \max \sum_{j \in J} c_{ij} x_{ij},$$

$$\sum_{j \in J} x_{ij} = 1, \tag{8.i}$$

$$0 \leqq x_{ij} \leqq y_j \quad \text{all } j \in J.$$

The optimal solution (6) follows. Furthermore, it is easy to check that $z_i(y)$ is a piecewise linear concave function of y. One way of doing it is by stating the dual of

the linear program (8.i). [This will also be useful in the proof of Proposition 2.]

$$z_i(y) = \min u_i + \sum_{j \in J} w_{ij} y_j,$$

$$(9.i) \qquad u_i + w_{ij} \geqq c_{ij} \quad \text{all } j \in J,$$

$$w_{ij} \geqq 0 \quad \text{all } j \in J.$$

To find an optimal solution of (9.i), it suffices to consider $u_i = c_{ik}$ for $k \in J$ and $w_{ij} = \max(0, c_{ij} - c_{ik})$, all $j \in J$.

Therefore

$$z_i(y) = \min_{k \in J} [c_{ik} + \sum_{j \in J} y_j \max(0, c_{ij} - c_{ik})].$$

$z_i(y)$ is the minimum of a finite number of linear functions and therefore it is piecewise linear and concave. The function $z_L(y) = \sum_{i \in I} z_i(y) - \sum_{j \in J} f_j y_j$ is the sum of piecewise linear concave functions and therefore is also piecewise linear and concave.

PROPOSITION 2. *Given a vector* $0 \leqq y \leqq 1$, *a subgradient of* z_L *at* y *is given by*

$$(10) \qquad v_j = \sum_{i \in I} \max(0, c_{ij} - c_{is_i(h_i)}) - f_j \quad \text{all } j \in J,$$

where h_i *is defined in* (7).

Proof. An optimal solution to (9.i) is

$$u_i^* = c_{is_i(h_1)} \quad \text{and} \quad w_{ij}^* = \max(0, c_{ij} - c_{is_i(h_i)}) \quad \text{all } j \in J.$$

Therefore

$$z_L(y) = \sum_{i \in I} z_i(y) - \sum_{j \in J} f_j y_j = \sum_{i \in I} u_i^* + \sum_{j \in J} \left(\sum_{i \in I} w_{ij}^* - f_j \right) y_j.$$

This equality identifies a linear function which achieves the minimum in the definition of z_L as the minimum of a finite number of linear forms. Therefore the coefficients of the variables y_j in this linear function form a subgradient of z_L at y. □

In fact, for this problem, the set of all the subgradients is easy to describe.

THEOREM 1. *The subdifferential of* z_L *at* y *is*

$$V = \{v(u) | v_j(u) = \sum_{i \in I} \max(0, c_{ij} - u_i) - f_j\},$$

where u *satisfies*

$$(11) \qquad c_{is_i(k_i)} \geqq u_i \geqq c_{is_i(l_i)} \quad \textit{for all } i \in I,$$

and k_i *and* l_i *are respectively the smallest and largest values of* h_i *satisfying* (7).

Proof. If u_i is outside the limits defined in (11), then setting u_i' equal to the violated bound yields a smaller value of $z_i(y)$ and therefore $z_L(y)$. Proposition 2 shows that each of the bounds in (11) defines a subgradient direction. Since V is defined as the convex hull of these bounds, the proof is complete. □

2. A primal subgradient algorithm. Given a concave real function to be maximized over a convex feasible set in R^n, a subgradient algorithm can be described by the following iterative process: calculate a subgradient of the function at a given point of the feasible set, update the point by making a step along the direction defined by

TABLE 1
Computational results.

Problem size	Fixed charge	# facilities opened at optimum	Optimum value	Value of greedy-interchange	Value of algorithm	CPU Time* (s) DEC-2060	
						Subgradient	Greedy-interchange
	5000	2	27474	27474	27474	3.6	0.2
	4000	2	25474	26861	25474	3.4	0.2
	3000	2	23474	23861	23627	3.3	0.3
	2500	3	22127	22361	22361	3.1	0.3
	2000	4	20363	20363	20363	2.8	0.2
33	1500	6	17832	17832	17832	2.3	0.2
	1000	6	14832	14832	14832	2.2	0.3
	500	10	11267	11287	11287	1.9	0.2
	295	17 or 18	8673	8691	8691	1.8	0.2
	184	31 or 32	6024	6024	6024	1.6	0.2
	5000	3	38547	38547	38617	10.5	1.2
	4000	3	35547	35617	35547	10.6	1.3
	3000	4	32136	32522	32136	9.1	1.4
57	2500	5	30022	30022	30022	8.6	1.6
	2000	6	27222	27222	27222	7.6	1.5
	1500	7	23943	23943	23943	6.7	1.1
	1000	9	20307	20425	20307	6.0	1.2
	500	13	15261	15324	15324	4.9	1.4
	200	29	9142	9189	9142	4.0	1.7
	50	55	2821	2821	2821	3.6	0.8
	8000	4	87889	87889	87889	29.5	6.3
	7000	5	83720	84106	83889	26.9	6.6
100	6000	5	78720	78720	78720	24.3	7.0
	5000	6	73073	73440	73073	23.1	7.1
	4000	7	66407	66440	66407	21.4	7.5
	3000	7	59407	60672	59407	18.8	7.1
	2900	8	58613	59672	58613	19.5	7.7
	2000	12	50103	50176	50103	16.5	8.9
	1150	16	38479	38505	38575	13.8	10.3
	1000	17	35965	36302	36031	13.4	9.9

* CPU time does not include I/O, nor cost-sorting time (for a very coarse sorting procedure, the average CPU time was:

$n = 33$ CPU = 0.2
57 1.6
100 11.0).

the subgradient and by projecting onto the feasible set, and repeat the process until a prespecified convergence criterion is satisfied.

The size of the steps is crucial to the success and the speed of the procedure. In [11] it is shown that a step size of a^k guarantees the convergence if $\sum_{k=0}^{\infty} a^k = +\infty$ and $\lim_{k\to\infty} a^k = 0$. However, the rate of convergence is slow. Practitioners have used successfully a step size a^k which is halved if no improvement of the function value is obtained during a given number s of consecutive iterations. Note that this choice of a^k violates the sufficient condition of [11]. Another possibility is simply to decrease the step size by a factor q at each iteration, again with no guarantee of convergence to an optimum. Other step sizes proposed in the literature [7], [11] require the knowledge of a *target value*, i.e., an upper estimate of the optimal value z^*. We tried several step sizes for our problem and concluded that, unless a sharp estimate of z^* can be obtained, the steps with a target value do not outperform the other choices of a^k. An upper estimate of z^* arises naturally from a feasible solution of the dual of (1)–(4); however, a sharp estimate requires a good dual solution, an approach contrary to the spirit of this paper—namely, to present a purely primal algorithm. Consequently, we settled for a step size a^k without target value.

For our application the convex feasible set is the n-dimensional hypercube $Z = \text{conv}\{0, 1\}^n$ onto which it is very easy to project. A more flexible procedure, delineated in [11] and applied to constrained optimization in the hypercube by Demjanov [4] allows one to overshoot the projection at each step by as much as the distance from the current point to its projection.

PROPOSITION 3. *Let v be a subgradient of z_L at point y. If $y_j = 1$ and $v_j \geqq 0$ (resp., $y_j = 0$ and $v_j \leqq 0$), let $v^a = (v_1, \cdots, v_{j-1}, a, v_{j+1}, \cdots, v_n)$ be a vector such that $-|v_j| \leqq a \leqq |v_j|$. Then the following inequalities hold.*

(i) $v^a \cdot (y' - y) \geqq v \cdot (y' - y) \geqq z(y') - z(y)$ *for all* $y' \in Z$,

(ii) $\cos(v, y' - y) \leqq \cos(v^a, y' - y)$.

Proof. If $y_j = 1$ and $v_j \geqq 0$, then $0 \leqq y'_j \leqq 1$ implies $-1 \leqq y'_j - y_j \leqq 0$ and $v_j(y'_j - y_j) \leqq a(y'_j - y_j)$, proving (i). Furthermore, $\|v^a\| \leqq \|v\|$, since $-|v_j| \leqq a \leqq |v_j|$. This, together with (i), proves (ii). The same reasoning holds when $y_j = 0$ and $v_j \leqq 0$. □

In practice we were not able to identify significant differences between various overshooting strategies and settled for the simple projection. Another improvement of the subgradient direction proposed by Camerini, Fratta and Maffioli [1] consists in taking a linear combination of the subgradient and the former direction of search. We tested this approach but abandoned it also as it calls for the use of an approximation of z^*.

An original feature of the approach that we propose is to incorporate the primal subgradient algorithm in a heuristic. The idea of this heuristic is that, after a limited number of iterations of the subgradient algorithm, certain coordinates of the current solution y^k have converged either to 0 or to 1. (The criterion that we use for convergence is described below in the algorithm.) These variables are then fixed definitively to 0 or to 1, thus reducing the size of the location problem to a set of unsettled variables. Depending on the size of this subproblem, an optimal solution can be attempted. In general, we propose the use of a well tested heuristic, such as greedy-interchange [2], for this final phase.

3. The algorithm.

Step 1 (Initialization). Choose an initial vector y^0 with $0 \leqq y_j^0 \leqq 1$, an initial step size a^0, a step reduction factor q and a maximum number of iterations kmax. Finally set the variable fixing criteria ε, α, Δy^0 and η^k for each iteration k. Set $k = 0$.

Step 2 (Subgradient). Find a subgradient v^k of z_L at y^k according to Proposition 2. If $y_j^k = 1$ and $v_j^k \geqq 0$ (resp., $y_j^k = 0$ and $v_j^k \leqq 0$), set $v_j^k = 0$.

Update: $y_j^{k+1} = y_j^k + a^k v_j^k / \|v^k\|$, all $j \in J$.

Set $a^{k+1} = a^k/q$ and increment k by 1.

If $y_j^k < 0$, set $y_j^k = 0$.

If $y_j^k > 1$, set $y_j^k = 1$.

Step 3 (Variable fixing heuristic).

$\delta y_j^k = |y_j^k - y_j^{k-1}|$ and $\Delta y_j^k = \max\{\delta y_j^k, \alpha \Delta y_j^{k-1}\}$.

If $\Delta y_j^k \leqq \varepsilon$ and $y_j^k \geqq 1 - \eta^k$, fix $y_j = 1$.

If $\Delta y_j^k \leqq \varepsilon$ and $y_j^k \leqq \eta^k$, fix $y_j = 0$.

Step 4 (Termination). If every variable y_j has been fixed, STOP. If not, and $k <$ kmax, go to Step 2. Finally if $k \geqq$ kmax and some variables are still unsettled, use a heuristic such as Greedy-Interchange to solve the remaining subproblem. STOP.

4. Computational results. Three sets of problems were implemented. They have as many potential location areas as consumer areas ($I = J$). The 33×33 and 57×57 problems first appeared in [8]; the 100×100 problems appeared in [9]. The revenue c_{ij} is set to the negative of the distance between i and j. All the fixed charges are equal. These problems have been solved in [2], [3], [5], and [14].

The following values were adopted:

$$\text{kmax} = 300, \qquad \varepsilon = 0.001,$$

$$a^0 = 1, \qquad \alpha = 0.9,$$

$$q = 1.03, \qquad \eta^k = .20\ k/\text{kmax}.$$

$$y_j^0 = 0.5 \text{ and } \Delta y_j^0 = 1 \quad \text{for all } j \in J,$$

The interchange heuristic proceeds in two stages: first with a set of n locations, where n is determined by the greedy heuristic, then with $n+1$ locations; the smaller value is retained.

Table 1 indicates the optimal value of each problem, the value obtained by a greedy-interchange only (set kmax = 1 in the algorithm), and finally the value yielded by the algorithm, consisting of a subgradient search followed by a greedy interchange heuristic. The computational time is divided in three components: initialization (sorting the cost matrix), subgradient search and greedy interchange heuristic.

Table 2 gives an idea of the quality of the subgradient component of the algorithm.

TABLE 2
Variable fixing heuristic.

Problem 100	8000	7000	6000	5000	4000	3000	2900	2000	1150	1000
# unsettled variables after 200 iterations	16	14	20	21	21	25	25	31	33	39
# unsettled variables after 300 iterations	10	12	16	18	18	19	18	26	31	33
# variables fixed to nonoptimal value after 300 iterations	0	1	0	0	0	0	0	0	Not known	Not known

REFERENCES

[1] P. M. CAMERINI, L. FRATTA AND F. MAFFIOLI, *On improving relaxation methods by modified gradient techniques*, Math. Programming, 3 (1975), pp. 26–34.

[2] G. CORNUEJOLS, M. L. FISHER AND G. L. NEMHAUSER, *Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms*, Management Sci., 23 (1977), pp. 789–810.

[3] G. CORNUEJOLS, *Analysis of algorithms for a class of location problems*, Tech. Rep. 382, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1978.

[4] V. F. DEM'JANOV, *Algorithms for some minimax problems*, J. Comput. Systems Sci., 2 (1968), pp. 342–380.

[5] D. ERLENKOTTER, *A dual-based procedure for uncapacitated facility location*, Oper. Res., 26 (1978), pp. 992–1009.

[6] A. M. GEOFFRION, *Lagrangean relaxation for integer programming*, Math. Programming Studies, 2 (1974), pp. 82–114.

[7] M. HELD, P. WOLFE AND H. P. CROWDER, *Validation of subgradient optimization*, Math. Programming, 6 (1974), pp. 62–88.

[8] R. L. KARG AND G. L. THOMPSON, *A heuristic approach to solving traveling salesman problems*, Management Sci., 10 (1964), pp. 225–248.

[9] P. KROLAK, W. FELTS AND G. MARBLE, *A man-machine approach towards solving the traveling salesman problem*, Comm. ACM, 14 (1971), pp. 327–334.

[10] S. C. NARULA, U. I. OGBU AND H. M. SAMUELSON, *An algorithm for the P-median problem*, Oper. Res., 25 (1977), pp. 709–713.

[11] B. T. POLYAK, *A general method of solving extremum problems*, Doklady Akad. Nauk SSSR, 174 (1967), =Soviet Math. Doklady, 8 (1967), pp. 593–597.

[12] ———, *Minimization of unsmooth functionals*, Ž Vyčisl. Mat. i Mat. Fiz, =USSR Computational Math. and Math. Phys., 9 (1969), pp. 509–521.

[13] C. S. REVELLE AND R. W. SWAIN, *Central facilities location*, Geographical Anal., 2 (1970), pp. 30–42.

[14] L. SCHRAGE, *Implicit representation of variable upper bounds in linear programming*, Math. Programming Studies, 4 (1975), pp. 118–132.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0012 $01.00/0

VERTICES BELONGING TO ALL OR TO NO MAXIMUM STABLE SETS OF A GRAPH*

P. L. HAMMER,† P. HANSEN‡ AND B. SIMEONE§

Abstract. The focus of the present paper is on the relations between the set D of optimal solutions of a maximum weighted stable set problem, and the set C of optimal solutions of its continuous relaxation. The main result is that if a variable takes a constant binary value in *all* $\hat{X} \in C$, then it takes the same value in *all* $X \in D$ (this may be contrasted with a well-known result of Nemhauser and Trotter, stating that if a variable takes a binary value in *some* $\hat{X} \in C$, then it takes the same value in *some* $X \in D$). For any graph G, the set P of the vertices j such that $\hat{X}_j$ has a constant binary value in all $\hat{X} \in C$, can be efficiently detected; moreover, the results in this paper imply that in the unweighted case, the subgraph induced by P has the "strong" König–Egerváry property and that the subgraph induced by the complement of P has a perfect 2-matching: actually, the maximum stable sets of G are in a 1-to-1 correspondence with those of the latter subgraph.

1. Introduction. Let $G = (V, E)$ denote[1] a finite, undirected graph without loops. For notational simplicity, we assume that $V = \{1, \cdots, n\}$. A *stable set* $S \subseteq V$ is a set of pairwise nonadjacent vertices, i.e., such that $(j, k) \notin E, \forall j, k \in S$. Let $c = (c_j)$ denote an n-vector ($n = |V|$) of positive *weights* given to the vertices of G. The *weight of a stable set* is defined as the sum of the weights of its vertices. We are interested in the problem of determining one or all *maximum weight stable sets* (MWSS) of G, and in the properties of these sets. It is well known that this problem may be expressed by the following integer linear program (denoted by ISP, for "integer stability problem"):

$$\max z = \sum_{j=1}^{n} c_j x_j$$

subject to

$$x_j + x_k \leqq 1 \qquad \forall (j, k) \in E,$$

$$x_j \in \{0, 1\}, \qquad j = 1, 2, \cdots, n.$$

To any optimal solution $X^* = (x_j^*)$ of the ISP there corresponds an MWSS S defined by $x_j^* = 1$ if $j \in S$ and $x_j^* = 0$ otherwise. Replacing the integrality constraints of ISP by

$$x_j \in [0, 1], \qquad j = 1, 2, \cdots, n,$$

yields the *continuous* relaxation of ISP (denoted by CSP). The relationship between the optimal solutions of the CSP and of the ISP has been studied by Nemhauser and Trotter [9], Picard and Queyranne [10], Pulleyblank [11] and Berge [3], [4]. The main result of Nemhauser and Trotter [9] is that if a variable x_j takes the value 1 (respectively 0) in an optimal solution $\hat{X}$ of CSP the corresponding vertex belongs (respectively

* Received by the editors November 3, 1981. This work was done while the first author was visiting professor at the Swiss Federal School of Technology, Lausanne, and, in part, during visits of the last two authors there and of the two first ones to the Istituto M. Picone, as well as while the second author was visiting professor at the Graduate School of Business, University of Pittsburgh. Support of the Swiss Federal School of Technology, the Canadian Natural Sciences and Engineering Research Council, the Italian Consiglio Nazionale delle Richerche and the Graduate School of Business, University of Pittsburgh is gratefully acknowledged.

† Department of Combinatorics and Optimization, Faculty of Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

‡ Institut d'Economie Scientifique et de Gestion, Lille, France, and Faculté Universitaire Catholique, Faculté des Sciences Economiques, Mons, Belgium.

§ Istituto M. Picone per le Applicazioni del Calcolo, Consiglio Nazionale delle Richérche, Rome, Italy.

[1] The definitions and notation are similar to those of Berge [2].

does not belong) to at least one MWSS of G. Picard and Queyranne [10] have shown that there is a unique maximal set of variables which are integral in the optimal solutions $\hat{X}$ of the CSP. Pulleyblank [11] has studied the unweighted graphs such that no variable is integer in any optimal solution to CSP—the *2-bicritical graphs*—and has shown that they form the overwhelming majority of random graphs. Another characterization of these graphs, in terms of *regularizable graphs*, is due to Berge [3], [4].

The main result of this paper is that if a variable takes a constant binary value in *all* optimal solutions of CSP, then it also takes the same value in *all* optimal solutions of ISP. Although it will be seen that the problem of determining *all* vertices belonging to all or to no MWSS of G is NP-complete, the set P of those variables taking a constant binary value in all optimal solutions of CSP can be detected in polynomial time. The graphs for which $P = \varnothing$ and those for which $P = V$ are characterized in § 3. Characterizations of P are presented in § 4. In the unweighted case, which is dealt with in § 5, the graphs for which $P = \varnothing$ are seen to be those having a perfect 2-matching while those for which $P = V$ are the graphs with the "strong" König–Egerváry property.

2. Persistency properties. It is well known that in every basic feasible solution $\hat{X}$ of CSP the variables $\hat{x}_j$ take only the values 1, 0 or $\frac{1}{2}$ (cf. Balinski [1]). We shall restrict our attention to feasible solutions having this property. Throughout this paper it is understood that "solutions" means "feasible solutions with $0, \frac{1}{2}, 1$ components". Accordingly, the set V may be partitioned a priori into 7 classes defined by the values taken by the $\hat{x}_j$ in all optimal solutions $\hat{X}$ of CSP:

$$V_1 : \{j \mid \forall \hat{X}, \hat{x}_j = 1\}, \qquad V_0 : \{j \mid \forall \hat{X}, \hat{x}_j = 0\},$$
$$V_{1/2} : \{j \mid \forall \hat{X}, \hat{x}_j = \tfrac{1}{2}\}, \qquad V_{1/2,1} : \{j \mid \forall \hat{X}, \hat{x}_j = 1 \text{ or } \hat{x}_j = \tfrac{1}{2}\},$$
$$V_{0,1/2} : \{j \mid \forall \hat{X}, \hat{x}_j = 0 \text{ or } \hat{x}_j = \tfrac{1}{2}\}, \qquad V_{0,1/2,1} : \{j \mid \forall \hat{X}, \hat{x}_j = 1 \text{ or } \hat{x}_j = 0 \text{ or } \hat{x}_j = \tfrac{1}{2}\},$$
$$V_{0,1} : \{j \mid \forall \hat{X}, \hat{x}_j = 1 \text{ or } \hat{x}_j = 0\}.$$

It is of course assumed that if j belongs to the class V, x_j takes each of the values feasible for V in at least one $\hat{X}$. We first show that $V_{0,1}$ is empty for all G, while this is not necessarily the case for the six other classes.

THEOREM 2.1. *If a variable x_j takes the value* 1 *in an optimal solution of the* CSP *and the value* 0 *in another optimal solution of the* CSP, *then there exists an optimal solution with components* $0, 1, \frac{1}{2}$ *of the* CSP *in which it takes the value* $\frac{1}{2}$.

Proof. Consider a graph G and two optimal solutions $\hat{X}_1$ and $\hat{X}_2$ of the corresponding CSP, such that $\hat{x}_{1j} = 1$ and $\hat{x}_{2j} = 0$. Let $K = \{k \mid \hat{x}_{1k} = 1 - \hat{x}_{2k}, k = 1, 2, \cdots, n\}$, $K_1 = \{k \mid k \in K, \hat{x}_{1k} = 1\}$ and $K_0 = \{k \mid k \in K, \hat{x}_{1k} = 0\}$.

Consider the n-vectors

$$X_3 = \begin{cases} \hat{x}_{1k}, & k \notin K, \\ \frac{1}{2}, & k \in K, \end{cases} \qquad X_4 = \begin{cases} \hat{x}_{2k}, & k \notin K, \\ \frac{1}{2}, & k \in K. \end{cases}$$

X_3 is feasible. Indeed, it is enough to prove that, if i and j are adjacent vertices and $x_{3i} = 1$, then one must necessarily have $x_{3j} = 0$. In fact, $x_{3i} = 1$ implies $x_{1i} = 1$ by the definition of X_3. But then $x_{1j} = 0$ since X_1 is feasible. Moreover, x_{2j} cannot be 1, otherwise x_{2i} would be 0 and thus i would belong to K and x_{3i} would be $\frac{1}{2}$. Hence we must have $x_{2j} < 1$, which implies $j \notin K$ and $x_{3j} = 0$. A similar argument shows that

X_4 is feasible. Now, we must have

$$\sum_{k\in K_1} c_k = \sum_{k\in K_0} c_k;$$

indeed,

$$\sum_{k\in K_1} c_k < \sum_{k\in K_0} c_k$$

would imply

$$z(X_3) = z(\hat{X}_1) - \tfrac{1}{2}\sum_{k\in K_1} c_k + \tfrac{1}{2}\sum_{k\in K_0} c_k > z(\hat{X}_1),$$

contradicting the optimality of $\hat{X}_1$, and a similar argument holds for the reverse inequality. But then X_3 and X_4 are optimal solutions also, proving the theorem. □

To show that the other classes V are nonempty for some G, we look at some unweighted graphs. The end vertices of a path of length 2 belong to V_1 and the middle vertex to V_0. The vertices of a triangle (or 3-cycle) belong to $V_{1/2}$ and those of a square (or 4-cycle) to $V_{0,1/2,1}$. In the graph of Fig. 1, the vertices 1 and 2 belong to $V_{1/2,1}$ and the vertices 3 and 4 to $V_{0,1/2}$.

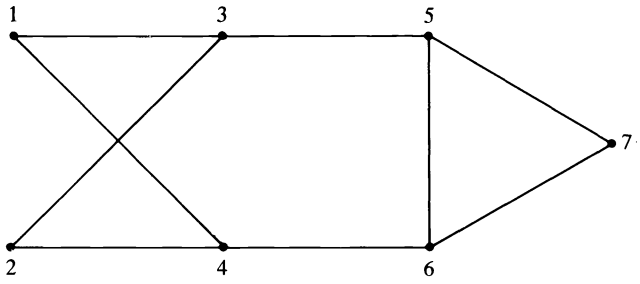


FIG. 1

An alternate proof of Theorem 2.1 could be obtained along the lines of the proof of Lemma 1 of Picard and Queyranne [10]. It is indeed implicit in that proof that the vector

$$X_5 = \begin{cases} 1, & \{k \mid \hat{x}_{1k} = \hat{x}_{2k} = 1\}, \\ 0, & \{k \mid \hat{x}_{1k} = \hat{x}_{2k} = 0\}, \\ \frac{1}{2}, & \{k \mid \hat{x}_{1k} \neq \hat{x}_{2k} \text{ or } x_{1k} = \frac{1}{2}\} \end{cases}$$

is an optimal solution to CSP; this immediately implies Theorem 2.1.

Note that our proof of Theorem 2.1 implies that the set of optimal solutions to CSP is closed under two binary operations, different from the three considered in Picard and Queyranne [10]. Using the notation of [8], these operations can be written as

$\circ$	0	$\frac{1}{2}$	1
0	0	0	$\frac{1}{2}$
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
1	$\frac{1}{2}$	1	1

and

$*$	0	$\frac{1}{2}$	1
0	0	$\frac{1}{2}$	$\frac{1}{2}$
$\frac{1}{2}$	0	$\frac{1}{2}$	1
1	$\frac{1}{2}$	$\frac{1}{2}$	1

.

We now prove a result stronger than Theorem 2.1.

THEOREM 2.2. *Let* $\bar{V} = V - V_0 - V_1$. *There is an optimal solution* $\hat{X}^*$ *of* CSP *such that* $\hat{x}_j^* = \frac{1}{2}$ *for all* $j \in \bar{V}$.

Proof. Let $\hat{X}_1$ be an optimal solution of CSP with a maximum number of components equal to $\frac{1}{2}$. Assume that the theorem is false; then there is an index $j \in \bar{V}$ such that $\hat{x}_{1j}$ is integer. Since $j \in \bar{V}$, there is an optimal solution $\hat{X}_2$ such that $\hat{x}_{1j} \neq \hat{x}_{2j}$. But then the vector X_5 defined above is an optimal solution to CSP with more components equal to $\frac{1}{2}$ than $\hat{X}_1$, a contradiction. □

Let us now turn to the relationship between the optimal values of the variables in ISP and CSP.

THEOREM 2.3. *If a variable* x_j *takes the value* 1 (*respectively* 0) *in all optimal solutions* $\hat{X}$ *of* CSP *then it retains the same value in all optimal solutions* X^* *of* ISP, *hence the corresponding vertex belongs to all* (*respectively to no*) *maximum weight stable sets of* G.

Proof. Assume by contradiction that the result does not hold. Then let $\hat{X}$ denote an optimal solution of CSP such that $\hat{x}_j$ is integer and X^* an optimal solution of ISP such that $x_j^* = 1 - \hat{x}_j$.

Let $K = \{k \mid \hat{x}_k = 0 \text{ or } \hat{x}_k = 1\}$. Consider the n-vector $\tilde{X}$ defined by

$$\tilde{x}_k = \begin{cases} \hat{x}_k & \text{for } k \in K, \\ x_k^* & \text{for } k \notin K. \end{cases}$$

$\tilde{X}$ is integer, and feasible. Indeed, for all (k, l) such that $k \in K, l \in K$ or such that $k \notin K, l \notin K$, this follows from the feasibility of $\hat{X}$ and of X^* respectively. For all (k, l) such that $k \in K, l \notin K$ and $(k, l) \in E$ we must have $\tilde{x}_k = \hat{x}_k = 0$ as $\hat{x}_l = \frac{1}{2}$.

Moreover,

$$\sum_{k \in K} c_k \hat{x}_k \leqq \sum_{k \in K} c_k x_k^*, \tag{2.1}$$

otherwise $z(\tilde{X}) > z(X^*)$, contradicting the optimality of X^*. But then, consider the n-vector, X' defined by

$$x'_k = \begin{cases} \frac{1}{2}\{\hat{x}_k + x_k^*\} & \text{for } k \in K, \\ \hat{x}_k & \text{for } k \notin K. \end{cases}$$

X' is feasible. This follows from the feasibility of $\hat{X}$ and X^* for $k \in K, l \in K$, the feasibility of $\hat{X}$ for $k \notin K, l \notin K$ and the fact that $\hat{x}_k = 0, \hat{x}_l = \frac{1}{2}$ for all (k, l) such that $k \in K, l \notin K$ and $(k, l) \in E$.

If (2.1) holds as a strict inequality $z(X') > z(\hat{X})$, contradicting the optimality of $\hat{X}$; if (2.1) holds as an equality X' is an optimal solution of CSP with $x'_j = \frac{1}{2}$, again a contradiction. □

The converse of Theorem 2.3 does not hold. Indeed, in the unweighted graph of Fig. 2, there is a single MSS consisting of 1 and 2, yet all vertices belong to $V_{1/2}$. No result similar to Theorem 2.3 holds if the vertices of G are assumed to belong to $V_{1/2,1}$ or $V_{0,1/2}$ instead of to V_1 or V_0. Indeed, the graph of Fig. 1 has four MSS: $\{1, 2, 5\}$, $\{1, 2, 6\}$, $\{1, 2, 7\}$ and $\{3, 4, 7\}$; so a vertex j such as 1, which is associated with a variable x_j never equal to 0 in an optimal solution of CSP, may not belong to all MSS of G, and a vertex j, such as 3, associated with a variable x_j never equal to 1 in an optimal solution of CSP, may belong to an MSS.

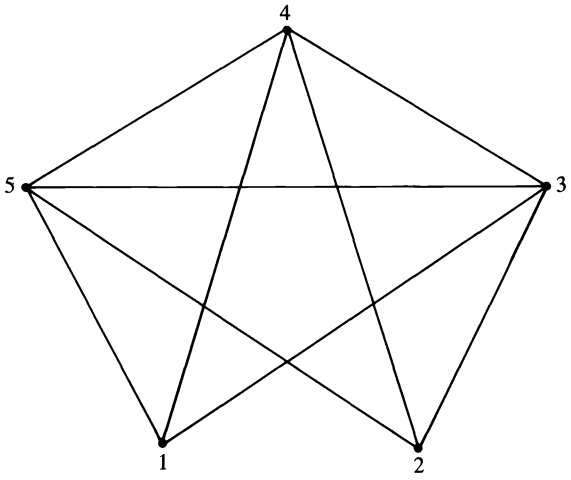


FIG. 2

The *permanent P* of G is the set of those vertices j for which x_j takes a constant binary (0, 1)-value in all optimal solutions of CSP. Determining the permanent can be done in polynomial time. Indeed, it is a well-known result of Edmonds and Pulleyblank (see Nemhauser and Trotter [9]) that an optimal solution of CSP can be determined by solving a maximum weight stable set problem in a bipartite graph $G'=(V', V'', E')$ associated with $G=(V, E)$; V' and V'' both have n vertices and $E'=\{\{i', j''\}, \{i'', j'\} | i, j \in E\}$, with weights $c'_j=c''_j=c_j$; associating with the optimal solution $\hat{X}', \hat{X}''$ of the weighted stability problem in G' the vector $\hat{X}=\frac{1}{2}(\hat{X}'+\hat{X}'')$, we note that $\hat{X}$ is an optimal solution to the CSP on G. To check if one of the conditions of Theorem 2.3 is satisfied, we consider in turn each variable x_j such that $\hat{x}_j=1$ or $\hat{x}_j=0$; we then set $x'_j=1, x''_j=0$ and resolve the bipartite weighted stable set problem; if the value of the objective function is $<z(\hat{X})$ the condition is satisfied, in view of Theorem 2.1.

As we have seen above, Theorem 2.3 allows us to easily detect, in general, only a subset of those variables belonging to all or to no MWSS of G. The next result shows that finding them all may be difficult, at least in some cases.

THEOREM 2.4. *The problem of determining all vertices belonging to all or to no maximum stable sets of a graph G (the persistent stable set problem) is* NP*-complete.*

Proof. We use Cook reducibility (see Cook [6], Garey and Johnson [7]). Assume that a polynomial algorithm exists to find all vertices belonging to all or to no maximum stable sets of G. Apply it to G and delete all such vertices. Each of the remaining vertices (if any) belongs to at least one maximum stable set of G. Choose any vertex and delete it, as well as its neighbors (i.e., adjacent vertices). Iterate this procedure until no more vertices remain. In this way, a maximum stable set would be determined in polynomial time. As the stability number problem is NP-complete, the persistent stable set problem is NP-hard.

To show that the persistent stable set problem is also NP-complete we consider all subgraphs G'_j and G''_j of G generated by V minus j and by V minus j and its neighbors respectively, for $j=1, 2, \cdots, n$. Now j belongs to all maximum stable sets of G if and only if $\alpha(G'_j)<\alpha(G)$; j belongs to no maximum stable set of G, if and only if $1+\alpha(G''_j)<\alpha(G)$. As the stability problem is NP-complete these conditions can be checked in polynomial time on a nondeterministic Turing machine, and the result follows. □

Theorem 2.5 allows us to further characterize the optimal solutions $\hat{X}$ of CSP. Let $\bar{V} = V - V_0 - V_1$ as above and G' denote the subgraph of G generated by $\bar{V}$.

THEOREM 2.5. (a) *The optimal solutions $\hat{X}$ of the* CSP *are precisely the vectors*

$$\hat{x}_j = \begin{cases} x_j^* & \text{if } j \in \bar{V}, \\ 1 & \text{if } j \in V_1, \\ 0 & \text{if } j \in V_0, \end{cases}$$

where X^* is an optimal solution of CSP *for G';*

(b) $X^{*\prime} = (\frac{1}{2}, \frac{1}{2}, \cdots, \frac{1}{2})$ *is an optimal solution of* CSP *for G'.*

Proof. To prove (a) let us first show that if $k \in \bar{V}$, $l \in V_0 \cup V_1$, and $(k, l) \in E$ then $\hat{x}_l = 0$ in all $\hat{X}$. Indeed, if the statement is false $\hat{x}_l = 1$ and hence $\hat{x}_k = 0$ in all $\hat{X}$, contradicting the assumption that $k \in \bar{V}$. Then (a) easily follows.

Assume (b) does not hold. Then by Theorem 2.2 there is some variable $\hat{x}_j$ which takes the same integer value in all optimal solution of CSP for G' and, because of (a), for G also. But this contradicts $j \in \bar{V}$. □

It follows from Theorem 2.5 that the process of finding those variables which take the same binary value in all optimal solutions of CSP for G, cannot be iterated on G'.

3. Extreme cases. In the present section, we shall give a characterization of "bad" graphs, for which *no* variable takes a constant binary value in all optimal solutions of the corresponding CSP, as well as of "good" graphs, for which *all* variables take a constant binary value in all optimal solutions of CSP. We denote by $\bar{z}$ the optimum value of CSP. If $S \subseteq V$, we set $c(S) = \sum_{j \in S} c_j$. We shall often refer to the three sets V_1, V_0, $\bar{V}$ defined above. If $S \subseteq V$, the *neighborhood* of S is the set $N(S) \equiv \{i | i \in V - S, (i, j) \in E \text{ for some } j \in S\}$. For simplicity, we write $N(i)$ instead of $N(\{i\})$ for singletons $\{i\}$.

A *perfect c-matching* of $G = (V, E)$ is a nonnegative vector $\lambda \in R^m$ $(m = |E|)$ such that

$$\sum_{j \in N(i)} \lambda_{ij} = c_i, \quad \text{for all } i = 1, \cdots, n.$$

Note that G may have no perfect c-matching.

A feasible solution of CSP will also be called a *fractional stable set* of G, while a nonnegative vector $\bar{X} = (\bar{x}_1, \cdots, \bar{x}_n)$ such that $\bar{x}_i + \bar{x}_j \geqq 1$ for all $(i, j) \in E$ will be called a *fractional transversal* of G. If, in addition, $\bar{X}$ is binary, then $\bar{X}$ is the characteristic vector of a *transversal* in the usual sense, i.e., a set C of vertices such that each edge has at least one endpoint in C.

THEOREM 3.1. *The permanent P of G is empty if and only if G has a perfect c-matching.*

Proof. According to Theorem 2.2, $P = \varnothing$ if and only if $h = (\frac{1}{2}, \cdots, \frac{1}{2})$ is a maximum fractional stable set of G; taking into account that $h = e - h$, where $e = (1, \cdots, 1)$, and the fact that for an arbitrary graph the set X is the characteristic vector of a maximum fractional stable set if and only if $e - X$ is the characteristic vector of a minimum fractional transversal, it follows that h is a minimum transversal of G.

Let us consider now the linear programming formulation of the minimum fractional transversal problem:

$$\begin{aligned} &\min \sum_{j=1}^{n} c_j \bar{x}_j \\ &\text{s.t. } \bar{x}_i + \bar{x}_j \geqq 1 \quad \text{for } (i, j) \in E, \\ &\bar{x}_j \geqq 0, \qquad j = 1, \cdots, n \end{aligned} \tag{3.1}$$

and its dual

$$\begin{aligned} &\max \sum_{(i,j)\in E} \lambda_{ij} \\ &\text{s.t.} \sum_{j\in N(i)} \lambda_{ij} \leqq c_i, \qquad i = 1, \cdots, n, \\ &\lambda_{ij} \geqq 0, \qquad (i,j)\in E. \end{aligned} \tag{3.2}$$

The vector $h = (\frac{1}{2}, \cdots, \frac{1}{2})$ is always a feasible solution of (3.1) with weight $(\sum_{j=1}^{n} c_j)/2$. On the other hand, by adding up all inequalities in (3.2)—except for the nonnegativity constraints—one gets $\sum_{(i,j)\in E} \lambda_{ij} \leqq (\sum_{j=1}^{n} c_j)/2$ for every feasible solution λ, the equality holding only when λ is a perfect c-matching. By linear programming duality it follows now that h is a minimum transversal of G if and only if G has a perfect c-matching. □

A graph $G=(V, E)$ will be said to be *c-tight* if V can be decomposed into two subsets U_1 and U_0 such that

(a) U_1 is stable,

(b) for all nonempty $U \subseteq U_0$, $c(U) < c(N(U) \cap U_1)$.

THEOREM 3.2. *A necessary and sufficient condition for all vertices of a graph G to belong to the permanent is that G is c-tight.*

Proof. Necessity. $U_1 = V_1$ and $U_0 = V_0$ are the required sets. V_1 is stable and for all nonempty $U \subseteq V_0$ one must have $c(U) < c(N(U) \cap U_1)$; otherwise the vector $\hat{X}$ with

$$\hat{x}_i = \begin{cases} 1 & \text{for } i \in V_1 - N(U), \\ 0 & \text{for } i \in V_0 - U, \\ \frac{1}{2} & \text{for } i \in U \cup (N(U) \cap V_1) \end{cases}$$

would be an optimal solution of CSP, a contradiction.

Sufficiency. Let us assume G to be c-tight and let us prove that the vector $\tilde{X}$ defined by

$$\tilde{x}_i = \begin{cases} 1 & \text{for } i \in U_1, \\ 0 & \text{for } i \in U_0 \end{cases}$$

is the unique optimal solution of CSP. In fact, let X be an arbitrary basic feasible solution of CSP and let

$$D_1 \equiv \{i | x_i = 1\}, \quad D_0 \equiv \{i | x_i = 0\}, \quad D_{1/2} \equiv \{i | x_i = \tfrac{1}{2}\}.$$

If $A = D_1 \cap U_0$, and $B = D_0 \cap U_1$, then $N(A) \cap U_1 \subseteq B$. By assumption, $c(A) < c(N(A) \cap U_1)$, unless $A = \varnothing$. Hence, the vector X' given by

$$x_i' = \begin{cases} 1, & i \in D_1 - A, \\ 0, & i \in D_0 - A, \\ \frac{1}{2}, & i \in D_{1/2} \cup A \cup (N(A) \cap U_1) \end{cases}$$

is feasible, and $cX' = cX + \frac{1}{2}(c(N(A) \cap U_1) - c(A)) \geqq cX$, the equality holding only when $D_1 = \varnothing$.

Next, let $F = (D_1 \cup D_{1/2}) \cap U_0$. Again by assumption, $c(F) < c(N(F) \cap U_1)$ unless $F = \varnothing$. Hence the vector X'' given by

$$x_i'' = \begin{cases} 1, & i \in D_1 \cup (N(F) \cap U_1), \\ 0, & i \in U_0, \\ \frac{1}{2}, & i \in D_{1/2} \cap U_1 \end{cases}$$

is feasible, and $cX'' = cX' + \frac{1}{2}(c(N(F) \cap U_1) - f(C)) \geqq cX'$, the equality holding if and only if $F = \varnothing$.

Finally, since $c\tilde{X} = cX'' + \frac{1}{2}c(D_{1/2} \cap U_1) + c(D_0 \cap U_1)$, one has $c\tilde{X} \geqq cX''$, with equality only if $U_1 = D_1$. It follows that $c\tilde{X} \geqq cX$, the equality holding only when $\tilde{X} = X$. Since X was arbitrary, $\tilde{X}$ is the unique optimal solution of CSP. Therefore $V_1 \subseteq U_1$ and $V_0 \subseteq U_0$.

Assume now that $\bar{V} \neq \varnothing$. We cannot have $\bar{V} \subseteq U_1$, since otherwise defining X^* by

$$x_i^* = \begin{cases} 1, & i \in V_1, \\ 0, & i \in V_0, \\ \frac{1}{2}, & i \in \bar{V}, \end{cases} \tag{3.3}$$

we would have $\bar{z} = c^T X^* < c^T \tilde{X} = \bar{z}$.

Hence the set $L = \bar{V} \cap U_0$ is not empty. Since V_1 and $\bar{V}$ are completely disconnected, $N(L) \cap U_1 \subseteq \bar{V}$. By hypothesis, $c(N(L) \cap U_1) > c(L)$. Hence the vector $\hat{X}$ given by

$$\hat{x}_i = \begin{cases} 1 & \text{for } i \in V_1 \cup (N(L) \cap U_1), \\ 0 & \text{for } i \in U_0, \\ \frac{1}{2}, & \text{for } i \in U_1 - V_1 - N(L) \end{cases}$$

is a feasible solution of CSP such that $c\hat{X} > cX^*$,a contradiction. It follows that $\bar{V} = \varnothing$. □

Remark. It is worth noting that the permanent of G coincides with V if and only if CSP has a unique optimal solution, and this solution is binary.

If $S \subseteq V$, we denote by $G(S)$ the subgraph of G induced by S.

For arbitrary graphs the following decomposition property holds.

THEOREM 3.3. *The vertex set of an arbitrary graph G can always be partitioned into two subsets P and Q such that $G(P)$ is c-tight and $G(Q)$ has a perfect c-matching.*

Proof. Let $P = V_1 \cup V_0$ be the permanent of G and $Q = \bar{V}$. By Theorems 2.5 and 3.1, $G(Q)$ has a perfect c-matching. On the other hand, V_1 is stable and for all $U \subseteq V_0$, we must have $c(U) < c(N(U) \cap V_1)$, for otherwise the vector $\tilde{X}$ given by

$$\tilde{x}_i = \begin{cases} 1 & \text{if } i \in V_1 - N(U), \\ 0 & \text{if } i \in V_0 - U, \\ \frac{1}{2} & \text{if } i \in \bar{V} \cup U \cup (N(U) \cap V_1) \end{cases}$$

would be a feasible solution of CSP such that $c^T \tilde{X} \geqq c^T X^* = \bar{z}$, with X^* given by (3.3). Hence $\tilde{X}$ would be an optimal solution for CSP, against the definition of V_1 and V_0. □

4. Characterization of the permanent. The results of this section give a characterization of the three sets V_1, V_0, $\bar{V}$ defined above.

Since $V_0 = N(V_1)$ and $\bar{V} = V - V_1 - N(V_1)$, the set V_1 uniquely determines V_0 and $\bar{V}$: hence we shall concentrate our attention only on V_1.

THEOREM 4.1. *A vertex belongs to V_1 if and only if it is unsaturated in some maximum c-matching.*

Proof. Let A be the set of those vertices j having the property that j is *unsaturated* in some maximum c-matching. Let us prove that $A \subseteq V_1$. In fact, if $i \in A$ there is a maximum c-matching λ such that $\sum_{j \in N(i)} \lambda_{ij} < c_i$. Applying complementary slackness to the primal-dual pair of linear programs (3.1) and (3.2), one sees that $\bar{x}_i = 0$ for

all optimal solutions $\bar{X}$ of (3.1), and thus $x_i = 1$ for all optimal solutions X of CSP. Hence $A \subseteq V_1$.

In order to prove that $V_1 \subseteq A$, consider an optimal solution $\bar{X}$ of (3.1) and an optimal solution $\bar{\lambda}$ of (3.2) such that $(\bar{X}, \bar{\lambda})$ is a strongly complementary pair. $X^* = e - \bar{X}$ is an optimal solution of CSP. If $i \in V_1$, one has $x_1^* = 1$, $\bar{x}_1 = 0$ and $\sum_{j \in N(i)} \lambda_{ij} < c_i$; hence $i \in A$. □

Another characterization of V_1 is based on the concept of "major set".

We shall call a stable set S^* a *major* set if S^* maximizes $c(S) - c(N(S))$ over all stable sets of G.

LEMMA 4.2. *A stable set S is major if and only if there is an optimal solution X of* CSP *such that $S \equiv \{i | x_i = 1\}$.*

Proof. Let S be an arbitrary stable set. Define a vector X by

$$x_i = \begin{cases} 1 & \text{if } i \in S, \\ 0 & \text{if } i \in N(S), \\ \frac{1}{2} & \text{if } i \in V - S - N(S). \end{cases}$$

X is a feasible solution of CSP, and

$$\bar{z} \geqq c^T X = \tfrac{1}{2}\left(\sum_j c_j\right) + \tfrac{1}{2} c(S) - c(N(S)). \tag{4.1}$$

On the other hand, if X^* is an optimal solution of CSP, let $S_1^* \equiv \{i : x_i^* = 1\}$ and $S_0^* \equiv \{i : x_i^* = 0\}$ (note that S_1^* and S_0^* may be empty). Then one must have $S_0^* = N(S_1^*)$. Hence

$$\bar{z} = c^T X^* = \tfrac{1}{2}\left(\sum_j c_j\right) + \tfrac{1}{2}(c(S_1^*) - c(N(S_1^*))). \tag{4.2}$$

The lemma then follows from (4.1) and (4.2). □

THEOREM 4.3. *For an arbitrary graph G there is a unique minimal major set, and this set is precisely V_1.*

Proof. The set V_1 is stable. Moreover, by Theorem 2.5, there is an optimal solution X^* of the CSP such that $V_1 \equiv \{i : x_i^* = 1\}$. Hence V_1 is a major set by Lemma 4.2. Moreover, again by Lemma 4.2 and by the definition of V_1, every major set contains V_1. Hence V_1 is the unique minimal major set of G. □

5. The unweighted case. For the unweighted case, i.e., the case when all the weights c_j are equal, the results of the previous sections take a particularly simple and interesting form.

For technical reasons, and without loss of generality, we may assume that $c_j = 2$ for $j = 1, \cdots, n$. Then, Theorem 3.1 implies

THEOREM 5.1. *A graph G has the property that no variable takes a constant binary value in all optimal solutions of the corresponding* CSP *if and only if the vertex set of G can be covered by pairwise nonincident edges and odd cycles.*

Proof. Follows from Theorem 3.1 and from the easily seen fact that G has a perfect 2-matching if and only if the vertex set of G can be covered by a set of pairwise nonincident edges and odd cycles. □

COROLLARY 5.2. *If the stability number $\alpha(G)$ of G is greater than $n/2$, then at least one vertex of G belongs to all maximum stable sets of G.*

Proof. If $\alpha(G)$ is greater than $n/2$, then also the fractional stability number $\alpha^*(G)$ is greater than $n/2$. Hence $h = (\frac{1}{2}, \cdots, \frac{1}{2})$ cannot be an optimal solution of CSP and thus, by Theorem 2.2, $V_0 \cup V_1$ must be nonempty. As a matter of fact, if V_0 is nonempty also V_1 must be such, for otherwise the vector $h = (\frac{1}{2}, \cdots, \frac{1}{2})$ would be an optimal solution of CSP. Thus one always has $V_1 \neq \varnothing$ and the statement then follows from Theorem 2.3. □

For example, in all bipartite graphs G with bipartition $\{A, B\}$ such that $|A| \neq |B|$, there exists at least one vertex belonging to all maximum stable sets of G.

Let us now turn our attention to those graphs for which the permanent coincides with the whole vertex set. We recall that a graph G is said to have the *König–Egerváry* (KE) *property* if, denoting by $\nu(G)$ the maximum cardinality of a matching and by $\tau(G)$ the minimum cardinality of a transversal of G, one has $\nu(G) = \tau(G)$.

Making use of P. Hall's theorem, it is not hard to see that the graphs with the KE property can be equivalently defined as those graphs G whose vertex set can be partitioned into two subsets U_1 and U_0 such that:

(5.1)
(a) U_1 is stable,
(b) For all $U \subseteq U_0$, $|N(U) \cap U_1| \geqq |U|$.

If strict inequality holds in (5.1) for all nonempty $U \subseteq U_0$, we shall say that G has the *strong König–Egerváry* (SKE) *property*.

If a graph G is such that $\bar{V} = \varnothing$, then G must necessarily have the KE property, since $\bar{V} = \varnothing$ implies $\alpha(G) = \alpha^*(G)$ and hence $\nu(G) = \tau(G)$ by a theorem of Lovász [8].

Theorem 3.2 then takes the form:

THEOREM 5.3. *The permanent of a graph G coincides with the vertex set of G if and only if G has the strong König–Egerváry property.*

If M is a matching, a vertex v is said to be *exposed* if no edge in M has v as an endpoint; otherwise v is said to be *saturated*. If v is saturated, the unique vertex u such that $(u, v) \in M$ is called the *twin* of v.

THEOREM 5.4. *A graph G has the strong König–Egerváry property if and only if*:

(a) *G has the König–Egerváry property,*

(b) *G has a unique minimum transversal C, and*

(c) *for all maximum matchings M, each vertex of C is connected to some exposed vertex by an alternating path.*

Proof. Necessity. Let G have the strong König–Egerváry property. Then (a) is obvious. (b) follows from Theorem 3.3 and from the remark that if CSP has a unique optimal solution X and if X happens to be binary, then also ISP must have X as its unique solution, and thus $\bar{X} = e - X$ yields the unique minimum transversal of G. In order to prove (c), let U_1 and U_0 be the two sets in the definition of the SKE property. As we have seen in the proof of Theorem 3.3, U_0 is a minimum transversal (and hence coincides with the unique minimum transversal C of G), and U_1 is a maximum stable set. If M is any maximum matching, each edge of M has exactly one endpoint in U_0. Let v_0 be a given vertex in U_0. Assign labels to the vertices of G as follows:

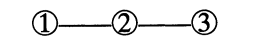
1. Give v_0 the label "−".
2. If v is an unscanned vertex labelled "−", give the label "+" to all unlabelled neighbors w of v in U_1. For every such w, set $p(w) = v$ (that is, v is declared to be the predecessor of w). Declare v scanned.
3. If u is an unscanned vertex labelled "+", and u is exposed, stop; in this case an alternating path from v_0 to u exists and can be retrieved from the pre-

decessor index $p(\cdot)$ by the usual backtracking procedure. Else if the twin w of u is unlabelled, give the label "$-$" to w, and set $p(w) = u$. Declare u scanned.

4. If there are no more unlabelled vertices, stop. In this case, let U be the set of vertices labelled "$-$" and let U' be the set of vertices labelled "$+$". Then clearly $|U'| = |U|$ and $U' = N(U) \cap U_1$, by the way the algorithm works. Hence G does not have the SKE property.

Sufficiency. If G has the KE property but does not have the SKE property, then there must be a nonempty set $U \subseteq U_0$ such that $|U| = |N(U) \cap U_1|$. Now for all maximum matchings M, each edge of M has exactly one endpoint in U_0. If v_0 is an arbitrary vertex of U then every neighbor of v_0 is saturated by M. Hence there cannot be any alternating path from v_0 to some exposed vertex. □

Let us call a set $S \subseteq V$ *bimatched* if it can be covered by pairwise nonincident edges and odd cycles. A vertex u will be called *avoidable* if there is a maximum bimatched set S such that $u \notin S$. For example, in the 2-path



there are just two avoidable vertices, namely 1 and 3.

THEOREM 5.5. *V_1 is the set of avoidable vertices of G.*

Proof. Consider the linear program

$$\begin{aligned} &\max \sum_{(i,j)\in E} \lambda_{ij} \\ &\text{s.t.} \sum_{j\in N(i)} \lambda_{ij} \leqq 2 \quad \text{for all } i = 1, \cdots, n, \\ &\lambda_{ij} \geqq 0 \qquad \text{for all } (i,j) \in E. \end{aligned} \tag{5.1}$$

A theorem of Tutte [12] states that if $\mathcal{M}$ is a collection of pairwise nonincident edges and odd cycles, then the vector λ defined by

$$\lambda_{ij} = \begin{cases} 2 & \text{for all edges } (i,j) \in \mathcal{M}, \\ 1 & \text{for all edges } (i,j) \text{ along odd cycles of } \mathcal{M}, \\ 0 & \text{for all other edges} \end{cases} \tag{5.2}$$

is a basic feasible solution of (5.1); and, conversely, all basic feasible solutions arise in this way.

If S is a bimatched set, $\mathcal{M}$ is a collection of pairwise nonincident edges and odd cycles covering S, and λ is defined through (5.2), then $\sum_{(i,j)\in E} \lambda_{ij} = |S|$. Hence if S is a maximum cardinality bimatched set then λ is a maximum 2-matching; conversely, if λ is a basic optimal solution of (5.1) then the set of saturated vertices is a maximum bimatched set.

It follows that a vertex i is avoidable if and only if it is unsaturated in some maximum 2-matching (note that if i is unsaturated in some maximum 2-matching, it is also unsaturated in some *basic* maximum 2-matching). Then the statement follows from Theorem 4.1. □

If we call a set $T \subseteq V$ *avoidable* when there is a maximum bimatched set S such that $S \cap T = \varnothing$, Theorem 5.4 can be rephrased by saying that V_1 is the union of all avoidable sets of G. Note that in Theorem 4.3, V_1 was characterized as the intersection of all major sets of G.

Our final result summarizes the relationships between the three sets V_1, V_0, $\bar{V}$ and the (discrete) maximum stable sets of G.

THEOREM 5.6. (a) *For all maximum stable sets S of G, $V_1 \subseteq S$ and $V_0 \cap S = \emptyset$;*
(b) *The maximum stable sets of G are precisely the sets $S = V_1 \cup S'$, where S' is a maximum stable set of $G(\bar{V})$.*

Proof. (a) follows at once from Theorem 2.5; (b) is an easy consequence of the relation $V_0 = N(V_1)$. □

In conclusion it appears that the determination of V_1, V_0 and $\bar{V}$, which can be carried out by a computationally efficient procedure, allows the reduction of a hard problem to another hard problem of smaller size and special structure.

Acknowledgments. The authors thank A. Rinnooy Kan and J. K. Lenstra for useful discussions.

REFERENCES

[1] M. L. BALINSKI, *Integer programming: methods, uses, computation*, Management Sci., 12 (1965), pp. 253–313.
[2] C. BERGE, *Graphes et hypergraphes*, Dunod, Paris, 1970; English translation: *Graphs and Hypergraphs*, North-Holland, Amsterdam, 1973.
[3] ———, *Regularisable graphs*, 1, Discrete Math., 23 (1978), pp. 85–89.
[4] ———, *Regularisable graphs*, 2, Discrete Math., 23 (1978), pp. 91–95.
[5] ———, *Packing problems and hypergraph theory: a survey*, Ann. Discr. Math., 4 (1979), pp. 3–38.
[6] S. A. COOK, *The complexity of theorem-proving procedures*, Proc. Third ACM Symposium on Theory of Computing, Association for Computing Machinery, New York, 1971, pp. 151–158.
[7] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability*, Freeman, San Francisco, 1979.
[8] L. LOVÁSZ, *Minimax theorems for hypergraphs*, in Hypergraph Seminar 1972, C. Berge and D. K. Ray-Chaudhuri, eds., Lecture Notes in Mathematics 411, Springer, Berlin, 1972, pp. 111–126.
[9] G. L. NEMHAUSER AND L. E. TROTTER, *Vertex packings: structural properties and algorithms*, Math. Programming, 8 (1975), pp. 232–248.
[10] J. C. PICARD AND M. QUEYRANNE, *On the integer valued variables in the linear vertex packing problem*, Math. Programming, 12 (1977), pp. 97–101.
[11] W. R. PULLEYBLANK, *Minimum node covers and 2-bicritical graphs*, Math. Programming, 17 (1979), pp. 91–103.
[12] W. T. TUTTE, *The factors of graphs*, Canadian J. Math., 4 (1952), pp. 314–328.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0013 $01.00/0

MINIMIZING A COMBINATORIAL FUNCTION*

DING ZHU DU† AND F. K. HWANG‡

Abstract. Let $M(N, d)$ denote the minimax number of group tests required for the identification of the d defectives in a set of N items. It is of interest to determine the values of N and d for which $M(N, d) = N-1$ (achieved by testing the first $N-1$ items one by one). Recently it has been proved that $M(N, d) = N-1$ for $N < \lfloor 2.5d \rfloor$. A lemma crucially used to obtain that result is the following:

$$M(N, d) \geqq \min\left\{N-1, 2t+\left\lceil \log_2 \binom{N-t}{d-t}\right\rceil\right\}.$$

The problem is to find a suitable t such that

$$N-1 \leqq 2t+\left\lceil \log_2 \binom{N-t}{d-t}\right\rceil$$

and d/N is minimized. However, standard methods do not work for this minimization problem. In this paper we propose a novel method to solve the minimization problem to obtain the new result: $M(N, d) = N-1$ for $N \leqq \lfloor 2.625d \rfloor$.

1. Introduction. Let m and n be relatively prime positive integers, l an integer satisfying $0 \leqq l \leqq m+n-2$, and λ a positive number. Define $l_1 = \lfloor m(l+1)/(m+n) \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer not exceeding x. The problem is to locate the minimum of

$$F(k) = \binom{(m+n)k+l}{mk+l_1} \lambda^k$$

over the nonnegative integers $k = 0, 1, 2, \cdots$.

Let $l_2 = \lfloor n(l+1)/(m+n) \rfloor$. Since m and n are relatively prime and $l+1 < m+n$, neither l_1 nor l_2 can be an integer. Therefore we have

$$\frac{m(l+1)}{m+n} - 1 < l_1 < \frac{m(l+1)}{m+n}, \qquad \frac{n(l+1)}{m+n} - 1 < l_2 < \frac{n(l+1)}{m+n}.$$

Adding up, we obtain

$$l-1 < l_1 + l_2 < l+1,$$

or $l_1 + l_2 = l$.

Define

$$f(x) = \frac{\lambda \prod_{i=0}^{m+n-1} [(m+n)(x+1)+l-i]}{\prod_{i=0}^{m-1} [m(x+1)+l_1-i] \prod_{i=0}^{n-1} [n(x+1)+l_2-i]}$$

for real $x \geqq 0$. Then

$$f(k) = \frac{F(k+1)}{F(k)} \quad \text{for } k = 0, 1, 2, \cdots.$$

If $f(x) = 1$ has no nonnegative solution, then $F(k)$ is either monotone increasing or monotone decreasing. If $f(x) = 1$ has a unique nonnegative solution x^0, then $F(k)$ has a minimum k^0 which is either $\lfloor x^0 \rfloor$ or $\lceil x^0 \rceil$, where $\lceil x \rceil$ denotes the smallest integer not less than x. Of course, since $f(x)$ is a polynomial of degree $m+n-1$, we would

* Received by the editors June 9, 1981.
† Institute of Applied Mathematics, Academy of Science, Beijing, China.
‡ Bell Laboratories, Murray Hill, New Jersey 07974.

not expect in general that $f(x)=1$ would have so few solutions. However, in this paper we give an interesting and novel method to show that indeed $f(x)=1$ has at most one nonnegative solution. We also show how this minimization problem can be applied to a group testing problem.

2. The main result. Two sets of numbers $A=\{a_1<a_2<\cdots<a_n\}$ and $B=\{b_1<b_2<\cdots<b_n\}$ are said to be *interleaved* if $b_1<a_1<b_2<a_2<\cdots<b_n<a_n$ (reversing A and B, if necessary).

LEMMA 1. *Suppose the set* $\{a_1<a_2<\cdots<a_n\}$ *and the set* $\{b_1<b_2<\cdots<b_n\}$ *are interleaved. Then the equation*

$$g(x)=\frac{\prod_{i=1}^{n}(x-a_i)}{\prod_{i=1}^{n}(x-b_i)}=c,$$

where $c>0$ *is a constant, has* $n-1$ *roots in the interval* (b_1, a_n), *and one root in the interval* $(S-\sum_{i=2}^{n} b_i, S-\sum_{i=1}^{n-1} a_i)$, *where* S *is defined by*

$$S=\frac{\sum_{i=1}^{n} a_i - c\sum_{i=1}^{n} b_i}{1-c}.$$

Proof. Since for $\varepsilon>0$ and $1\leqq i\leqq n$

$$g(b_i-\varepsilon)\xrightarrow[\varepsilon\to 0]{}\infty \quad \text{and} \quad g(a_i)=0,$$

there exists at least one root in the interval (a_i, b_{i+1}) for each $i=1, 2, \cdots, n-1$ due to the continuity of $g(x)$ in that interval. Let $x_1, x_2, \cdots, x_n$ denote the n roots. Then by comparing the coefficients of both sides of the equation

$$\prod_{i=1}^{n}(x-a_i)-c\prod_{i=1}^{n}(x-b_i)=(1-c)\prod_{i=1}^{n}(x-x_i)$$

we have

$$\sum_{i=1}^{n} x_i=\frac{\sum_{i=1}^{n} a_i - c\sum_{i=1}^{n} b_i}{1-c}=S.$$

By noting that the sum of the previously found $n-1$ roots is greater than $\sum_{i=1}^{n-1} a_i$ but less than $\sum_{i=2}^{n} b_i$, Lemma 1 follows immediately.

Let I denote the set $\{(l+i)/(m+n), i=1, 2, \cdots, m+n, i\neq m+n-l\}$, J the set $\{(l_1+j)/m, j=1, 2, \cdots, m, j\neq m-l_1\}$ and K the set $\{(l_2+k)/n, k=1, 2, \cdots, n, k\neq n-l_2\}$.

LEMMA 2. *Suppose that* m *and* n *are relatively prime. Then the set* $J\cup K\cup\{1\}$ *and the set* I *are interleaved.*

Proof. We first observe that no number in $J\cup K\cup\{1\}$ can be equal to a number in I. For example, if

$$\frac{l+i}{m+n}=\frac{l_1+j}{m} \quad \text{for some } 1\leqq i\leqq m+n \text{ and } 1\leqq j\leqq m,$$

then

$$\frac{m}{n}=\frac{l_1+j}{l_2+i-j}.$$

Since m and n are relatively prime and $l_1+j<2m$, necessarily $l_1+j=m$ and $l+i=m+n$. Hence i and j are not in the designated sets.

Next we show that there exists at most one number in $J \cup K$ contained in the interval $((l+i)/(m+n), (l+i+1)/(m+n))$ for each $i = 1, 2, \cdots, m+n, i \neq m+n-l$. Suppose to the contrary that there exist two numbers in $J \cup K$ contained in the interval $((l+i)/(m+n), (l+i+1)/(m+n))$. Since $1/m > 1/(m+n)$ and $1/n > 1/(m+n)$, the two numbers cannot be both in J or both in K. So we can let $(l_1+j)/m$ and $(l_2+k)/n$ denote the two numbers. Then

$$\frac{l+i}{m+n} < \frac{l_1+j}{m} < \frac{l+i+1}{m+n}, \qquad \frac{l+i}{m+n} < \frac{l_2+k}{n} < \frac{l+i+1}{m+n},$$

or equivalently

$$\frac{l_1+j}{l+i+1} < \frac{m}{m+n} < \frac{l_1+j}{l+i}, \qquad \frac{l_2+k}{l+i+1} < \frac{n}{m+n} < \frac{l_2+k}{l+i}.$$

But this implies

$$\frac{l+j+k}{l+i+1} < \frac{m+n}{m+n} < \frac{l+j+k}{l+i},$$

or $i < j+k < i+1$, a contradiction to the fact that i, j, k are all integers.

Finally, we observe that the number 1 must lie in the interval $((m+n-1)/(m+n), (m+n+1)/(m+n))$.

Define $M = \max\{(m+l_1)/m, (n+l_2)/n\}$. Since all numbers except M in $J \cup K \cup \{1\}$ lie in the interval $(1/(m+n), (m+n+l)/(m+n))$, and since I and $J \cup K \cup \{1\}$ have the same cardinality, Lemma 2 follows immediately.

COROLLARY. *Without loss of generality, assume* $M = (m+l_1)/m$. *Then Lemma 2 remains true if we decrease* l_1 *and increase* l_2 *each by* 1.

Define

$$c = \frac{m^m n^n}{(m+n)^{m+n}\lambda}.$$

THEOREM 1. *If* $c > 1$ *or* $c \leqq 2(l+1)/(m+n+2(l+1))$, $f(x) = 1$ *has no nonnegative solution. If* $1 \geqq c > 1 - 1/2M$, $f(x) = 1$ *has a unique nonnegative solution, lying in the interval* $(1/2(1-c) - M, c/2(1-c) - (l+1)/(m+n))$. *If* $1 - 1/2M \geqq c > 2(l+1)/m + n + 2(l+1)$, $f(x) = 1$ *either has no nonnegative solution or has a unique one, lying in the interval* $[0, c/2(1-c) - (l+1)/(m+n))$.

Proof. Note that $f(x)$ can be written as

$$f(x) = \frac{\lambda (m+n)^{m+n} \prod_{\substack{i=l+1 \\ i \neq m+n}}^{m+n+l} \left(x + \frac{i}{m+n}\right)}{m^n n^n (x+1) \prod_{\substack{i=l_1+1 \\ i \neq m}}^{m+l_1} \left(x + \frac{i}{m}\right) \prod_{\substack{i=l_2+1 \\ i \neq n}}^{n+l_2} \left(x + \frac{i}{n}\right)}.$$

Let

$$I' = \left\{ -\frac{l+i}{m+n}, i = 1, 2, \cdots, m+n, i \neq m+n-l \right\},$$

$$J' = \left\{ -\frac{l_1+j}{m}, j = 1, 2, \cdots, m, j \neq m-l_1 \right\},$$

$$K' = \left\{ -\frac{l_2+k}{n}, k = 1, 2, \cdots, n, k \neq n-l_2 \right\}.$$

From Lemma 2, the set $J' \cup K' \cup \{-1\}$ and the set I' are interleaved, since I', J' and K' are simply the negatives of I, J, and K, respectively. Define a_i and b_i for $i = 1, 2, \cdots, m+n-1$, such that

$$A = \{a_1 < a_2 < \cdots < a_{m+n-1}\} = I',$$

$$B = \{b_1 < b_2 < \cdots < b_{m+n-1}\} = J' \cup K' \cup \{-1\}.$$

Then $f(x) = 1$ implies

$$\frac{\prod_{i=1}^{m+n-1}(x-a_i)}{\prod_{i=1}^{m+n-1}(x-b_i)} = c.$$

From Lemma 1, $f(x) = 1$ has $m+n-2$ roots lying in the interval $(-M, -(l+1)/(m+n))$ and one root in the interval $(1/2(1-c)-M, c/2(1-c)-(l+1)/(m+n))$. Then we have

$$\sum_{i=1}^{m+n-1} a_i = 1 - \sum_{i=1}^{m+n} \frac{l+i}{m+n} = 1 - l - \frac{m+n+1}{2},$$

$$\sum_{i=1}^{m+n-1} b_i = 1 - \sum_{j=1}^{m} \frac{l_1+j}{m} - \sum_{k=1}^{n} \frac{l_2+k}{n} = 1 - l - \frac{m+n+2}{2},$$

$$S = \frac{\sum_{i=1}^{m+n-1} a_i - c\sum_{i=1}^{m+n-1} b_i}{1-c} = 1 - l - \frac{m+n+1}{2} + \frac{c}{2(1-c)},$$

$$S - \sum_{i=1}^{m+n-2} a_i = \frac{c}{2(1-c)} - \frac{l+1}{m+n},$$

$$S - \sum_{i=2}^{m+n-1} b_i = \frac{1}{2(1-c)} - M.$$

Since the last interval is the only one that could contain a nonnegative root, Theorem 1 follows immediately.

COROLLARY. *If* $c > 1$, $F(k)$ *is monotone decreasing. If* $1 \geqq c > 2(l+1)/(m+n+2(l+1))$, *then* $F(k-1) > F(k)$ *for* $k \leqq \max\{0, \lceil 1/2(1-c) - M \rceil\}$ *and* $F(k+1) > F(k)$ *for* $k \geqq \lceil c/2(1-c) - (l+1)/(m+n) \rceil$ (*since* $c/2(1-c) - (l+1)/(m+n) - 1/2(1-c) + M < 1$, *we can obtain the minimum of* $F(k)$ *by comparing at most two values of* k *from* $\lceil 1/2(1-c) - M \rceil$ *to* $\lceil c/2(1-c) - \lceil (l+1)/(m+n) \rceil \rceil$). *If* $2(l+1)/(m+n+2(l+1)) \geqq c$, $F(k)$ *is monotone increasing.*

Proof. Obvious from the observation that

$$\frac{F(k+1)}{F(k)} \xrightarrow[k \to \infty]{} \frac{1}{c}.$$

3. An application to group testing. In a group testing problem we have a set of N items including d defectives and $N-d$ good items. The problem is to identify all the defectives by means of a sequence of group tests where a group test is a simultaneous test on an arbitrary subset of items with two possible outcomes. A "pure" outcome indicates that every item in the subset is good while a "contaminated" outcome indicates that at least one item in the subset is defective. For given N and d, let $M_T(N, d)$ denote the worst-case number of tests required by the group testing algorithm T to solve the (N, d) problem. Define

$$M(N, d) = \min_T M(N, d).$$

Let I denote the algorithm which tests each item individually. Then clearly

$$M_I(N, d) = N - 1,$$

since the nature of the last item can be deduced from knowledge of the other items, while no other items can be exempted from testing in the worst case. The question arises: For what values of N and d, is it the case that $M(N, d) = M_I(N, d) = N - 1$? Recently Hu, Hwang and Wang [1] proved that

$$M(N, d) = N - 1 \quad \text{for } 2N \leqq 5d + 1$$

and

$$M(N, d) < N - 1 \quad \text{for } N \geqq 3d$$

by using the following two lemmas.

LEMMA 3.

$$M(N, d) \geqq \min\left\{N - 1, 2t + \left\lceil \log_2 \binom{N-t}{d-t} \right\rceil\right\}$$

for $N > d \geqq t > 0$.

LEMMA 4. $M(N, d) = N - 1$ *implies* $M(N', d) = N' - 1$ *for* $d < N' < N$.

We now show that Theorem 1 and these two lemmas lead to a stronger result.

THEOREM 2. $M(N, d) = N - 1$ *for* $N \leqq \frac{21}{8}d$.

Proof. From Lemma 4, it suffices to prove Theorem 2 for $N = \lfloor \frac{21}{8}d \rfloor$.

We decompose the proof into eight cases.

Case (i). $d = 8k$. Then $N = \lfloor \frac{21}{8}d \rfloor = 21k$. Set $t = 4k$. Then $n - t = 17k$. $d - t = 4k$, $n - 2t - 2 = 13k - 2$. We will prove $\binom{17k}{4k} > 2^{13k-2}$ by showing that

$$\min_k \binom{17k}{4k} \Big/ 2^{13k-2} > 1.$$

Theorem 2 then follows from Lemma 3.

Define

$$F(k) = \binom{17k}{4k} \Big/ 2^{13k}$$

Then we have $m = 4$, $n = 13$, $l = l_1 = 0$, $\lambda = 2^{-13}$. We compute

$$M = \max\left(\frac{m + l_1}{m}, \frac{n + l_2}{n}\right) = 1,$$

$$1 > c = \frac{m^n n^n}{(m+n)^{m+n}\lambda} = .7677 > 1 - \frac{M}{2} = .5.$$

Therefore, from Theorem 1, $f(k) = 1$ has a unique nonnegative solution in the interval (1.15, 2.59). Namely, $F(k)$, and hence $\binom{17k}{4k}/2^{13k-2}$ attains a minimum at $k = 2$. Thus we have

$$\min_k \binom{17k}{4k} \Big/ 2^{13k-2} = \binom{34}{8} \Big/ 2^{24} = 1.08 > 1.$$

As the proofs for the other seven cases are identical to case (i) with different parameter values, we only give the values of the parameters in each case without further details. K^0 will always denote the value of k that minimizes $\binom{N-t}{d-t}/2^{N-2t-2}$.

Case (ii). $d = 8k+1$, $N = 21k+2$, $t = 4k+1$, $1.08 < K^0 < 2.54$.

$$\min_k \binom{17k+1}{4k} \Big/ 2^{13k-2} = \binom{35}{8} \Big/ 2^{24} = 1.40 > 1.$$

Case (iii). $d = 8k+2$, $N = 21k+5$, $t = 4k+2$, $.92 < K^0 < 2.42$.

$$\min_k \binom{17k+3}{4k} \Big/ 2^{13k-1} = \min\left\{\binom{20}{4} \Big/ 2^{12} = 1.18, \binom{37}{8} \Big/ 2^{25} = 1.15\right\} > 1.$$

Case (iv). $d = 8k+3$, $N = 21k+7$, $t = 4k+2$, $.84 < K^0 < 2.30$.

$$\min_k \binom{17k+5}{4k+1} \Big/ 2^{13k+1} = \min\left\{\binom{22}{5} \Big/ 2^{14} = 1.60, \binom{39}{9} \Big/ 2^{27} = 1.58\right\} > 1.$$

Case (v). $d = 8k+4$, $N = 21k+10$, $t = 4k+3$, $.69 < K^0 < 2.18$.

$$\min_k \binom{17k+7}{4k+1} \Big/ 2^{13k+2} = \min\left\{\binom{24}{5} \Big/ 2^{15} = 1.29, \binom{41}{9} \Big/ 2^{28} = 1.30\right\} > 1.$$

Case (vi). $d = 8k+5$, $N = 21k+13$, $t = 4k+3$, $.54 < K^0 < 2.01$.

$$\min_k \binom{17k+10}{4k+2} \Big/ 2^{13k+5} = \min\left\{\binom{27}{6} \Big/ 2^{18} = 1.13, \binom{44}{10} \Big/ 2^{31} = 1.16\right\} > 1.$$

Case (vii). $d = 8k+6$, $N = 21k+15$, $t = 4k+3$, $.40 < K^0 < 1.88$.

$$\min_k \binom{17k+12}{4k+3} \Big/ 2^{13k+7} = \binom{29}{7} \Big/ 2^{20} = 1.48 > 1.$$

Case (viii). $d = 8k+7$, $N = 21k+18$, $t = 4k+4$, $.31 < K^0 < 1.77$.

$$\min_k \binom{17k+14}{4k+3} \Big/ 2^{13k+8} = \binom{31}{7} \Big/ 2^{21} = 1.25 > 1.$$

REFERENCE

[1] M. C. HU, F. K. HWANG AND J. K. WANG, *A boundary problem for group testing*, this Journal, 2 (1981), pp. 81–87.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0014 $01.00/0

FIXED POINT BEHAVIOR OF THRESHOLD FUNCTIONS ON A FINITE SET*

ERIC GOLES†

Abstract. In this paper we obtain a sufficient condition that a kind of iteration scheme has no cycles other than fixed points.

A detailed version of this result and of its applications may be found in E. Goles [Tech. Rep., Depto. Matem., Univ. de Chile, Santiago, 1981].

1. Introduction. Given a real symmetric $(n \times n)$-matrix (a_{ij}), a real n-component threshold vector (t_i), and a partition of n into $n_1, \cdots, n_k$, we define a corresponding mapping M from the set of binary n-vectors into itself, as follows:

For indices up to the n_1th we set

$$M_i(\bar{V}) = 1 \quad \text{iff} \quad \sum a_{ij} V_j \geqq t_i.$$

For indices between $n_1 + 1$ and $n_1 + n_2$ we make the analogous computation, using however the value $M_j(\bar{V})$ instead of V_j for the first n_1 values:

$$M_i(\bar{V}) = 1 \quad \text{iff} \quad \sum_{j \leqq n_1} a_{ij} M_j + \sum_{j > n_1} a_{ij} V_j \geqq t_i.$$

In general, for indices lying in the kth block of indices, we use the M_j values rather than the original V_j, on all indices in the first $k-1$ blocks. This use of the most recent values corresponds to a Gauss–Seidel rather than a Jacobi iteration scheme.

For

$$S_k \equiv \sum_{j=1}^{k-1} n_j < i < \sum_{j=1}^{k} n_j$$

we have

$$M_i(\bar{V}) = 1 \quad \text{iff} \quad F_i(\bar{V}) \equiv \sum_{j \leqq s_k} a_{ij} M_j + \sum_{j > s_k} a_{ij} V_j \geqq t_i.$$

If $n_1 = n$, this is the Jacobi iteration scheme; all new values are computed using the old input value. For $n_i = 1$ it corresponds to the Gauss–Seidel iteration scheme where each new value is used as soon as it is computed. The general case could be called a *block Gauss–Seidel* scheme; numbers of values are computed at each stage with the same inputs, and all are updated together in determining the next set of components.

This kind of iteration arises in a number of contexts. An application of the results below is a study of the behavior of the spin systems and will be described.

2. The main result. The main purpose of this paper is to give a sufficient condition that the mapping $\bar{V} \to M(\bar{V})$ have no nontrivial cycles. Our main result is:

THEOREM. *If for every index i, a_{ii} is at least as large as the sum of the magnitudes of $|a_{ij}|$ for j in the same block as i but not equal to i, then $\bar{V} \to M(\bar{V})$ has only fixed points as cycles.*

Proof. If $n_1 = n$, so that there is only one block, the condition of the theorem implies that for any $\bar{V}$ with $V_i = 1$, $\sum a_{ij} V_j$ is no less than $\sum a_{ij} W_j$ for any vector $\bar{W}$ having $W_i = 0$. Thus if $\sum a_{ij} W_j \geqq t_i$, so that $M(\bar{W})_i = 1$, then we must also have $M^j(\bar{W})_i = 1$ for any $j \geqq 1$.

* Received by the editors July 25, 1981, and in revised form January 22, 1982.

† IMAG, B.P. 53X, 38041 Grenoble Cedex, France.

In general, the condition of the theorem implies that the contribution to $F_i(\bar{V})$ from indices within the same block as i (between $S+1$ and $S+n_k$) are larger if $V_i=1$ than the corresponding contributions to $F_i(\bar{W})$, if $W_i=0$.

The proof is completed by showing that the interblock contributions to $F_i(\bar{V})$ cannot be such as to permit cycling. Suppose, on the contrary, there were a cycle consisting of vectors $\bar{V}(1), \bar{V}(2), \cdots, \bar{V}(q), \bar{V}(1)=\bar{V}(q+1)$ such that

$$\bar{V}(j+1)=M(\bar{V}(j)).$$

Any index that takes the same value in each $\bar{V}(r)$ contributes identically to each $F_i(\bar{V}(n))$ and, therefore, will not contribute to differences among them. If the ith component of $\bar{V}(i)$ goes from zero to one at some point in the cycle and from one back to zero at some other point, the value of $F_i(\bar{V}(r))$ at the latter point must be strictly less than its value at the first one. This condition implies that

$$\sum_{j=1}^{q} (V_i(j))(F_i(\bar{V}(j))-F_i(\bar{V}(j-1)))<0.$$

The sum over any consecutive block of 1's for $V_i(j)$ telescopes to form the difference just discussed between the values at the point where $\bar{V}$ goes to zero and the point where it goes to one. The sum indicated is the sum of the strictly negative contributions over the various consecutive blocks of ones in the ith component, over the cycle.

We have already seen that the contributions from within i's block to this difference must be nonnegative; the interblock contributions must, therefore, be negative definite. Explicitly, they are

$$\sum_{j\leqq S} a_{ij} \sum_{p=1}^{q} V_i(p)(V_j(p+1)-V_j(p))$$

$$+\sum_{j>S+n_k} a_{ij} \sum_{p=1}^{q} V_i(p)(V_j(p)-V_j(p-1))<0.$$

If we sum this over all indices i, the contribution from indices for which $V_i(P)$ is independent of P trivially vanishes so that, if there is a nontrivial cycle, the entire sum must be strictly negative. However, for every pair of indices (i,j) with $i<j$ in different blocks, we obtain

$$a_{ij} \sum_{p=1}^{q} (V_i(p)V_j(p+1)-V_i(p)V_j(p))$$

from the ith term, and $a_{ij}\sum_{p=1}^{q} V_j(p)(V_i(p)-V_i(p-1))$ from the jth term. These contributions are equal in magnitude and opposite in sign, so that the entire sum is zero, and there cannot be any indices that vary over the cycle.

It is easily seen that the theorem is not valid in the nonsymmetrical case [4].

3. Application. Our theorem can be illustrated by a collection of *spins* or 0, 1 variables located at a rectangular array of lattice points in the plane. A state of this system is characterized by assigning 0 or 1 to each spin. The interactions among these spins are such that a spin will be influenced to flip its value by the orientations of its neighbors. For example, one can consider a model in which the value of a spin at time t_k is obtained by "majority vote" among its nearest neighbors (with no change in case of ties).

There are several versions of this model, depending upon which spin values are used for the neighbors when taking the vote to determine a particular spin at time t_k.

If the t_{k-1} spins are used, the system can be described by a threshold matrix model of the Jacobi kind already discussed—with $n_1 = n$. If one instead iterates row by row, using the t_k values for the previously considered rows in computing each row vote, one has a block Gauss–Seidel iteration with a block for each row.

One can also iterate in the Gauss–Seidel manner ($n_i = 1$) using all previously determined t_k spins in computing each t_k spin.

Our theorem implies that the majority rule, with no change in case of a tie, cannot give rise to oscillation in the last case, but can (and in fact often will) give rise to oscillation in the other cases. In the row by row case, the effect of the "inertial" diagonal term has to exceed that of all others in the same row to avoid oscillation. One can in fact have cycles with long periods if the conditions of the theorem are not satisfied in either of these cases.

Acknowledgments. I am indebted to the referee for careful reading of an earlier version of this paper.

REFERENCES

[1] ANGLES D'AURIAC AND P. VILLON, *Fluctuations d'aimantation dans un verre de spin par simulation numérique de Monte Carlo*, Rapport DEA, Analyse Numerique, Grenoble, France, 1978.

[2] E. GOLES, *Comportement oscillatoire d'une famille d'automates cellulaires uniformes*, These Docteur-Ingenieur, IMAG, Grenoble, France, 1980.

[3] E. GOLES AND J. OLIVOS, *Comportement périodique des fonctions à seuil binaires et applications*, Discrete Applied Math., 3 (1981), pp. 93–105.

[4] E. GOLES, *Fixed point behaviour of threshold functions*, Technical Report, Depto. Matem. U. de Chile, Santiago, 1981.

[5] F. ROBERT, *Comparaison des modes opératoires d'un automate cellulaire fini*, R.R. 31, IMAG, Grenoble, France, 1976.

[6] ———, *Une approche booléenne du problème de la frustration*, S.A.N.G. 302, IMAG, Grenoble, France, 1978.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0015 $01.00/0

SPEED-UP IN DYNAMIC PROGRAMMING*

F. FRANCES YAO†

Abstract. Dynamic programming is a general problem-solving method that has been used widely in many disciplines, including computer science. In this paper we present some recent results in the design of efficient dynamic programming algorithms. These results illustrate two approaches for achieving efficiency: the first by developing general techniques that are applicable to a broad class of problems, and the second by inventing clever algorithms that take advantage of individual situations.

1. Introduction. Dynamic programming is a general problem-solving technique that has been used widely in operations research, economics, control theory and, more recently, computer science. The present paper will be oriented toward the use of dynamic programming as a paradigm for designing algorithms in computer science. As computational efficiency is a major goal in algorithm design, we will be interested in techniques which allow us to speed up algorithms produced by straightforward dynamic programming. There are two promising directions for such research, namely, the development of general techniques that are applicable to a large class of problems, and the invention of efficient algorithms for specific problems by taking advantage of their special properties. In this paper we give a review of some recent progress in these directions. In § 2, we discuss a general speed-up technique that can be applied to dynamic programming problems when the cost function satisfies certain restrictions known as the Quadrangle Inequalities. In § 3, we give an improved algorithm for finding the optimal order of multiplying a sequence of matrices.

We will not give proofs for the theorems cited in this paper. For proofs as well as further discussions, the reader is referred to [8] for the topic considered in § 2, and [4], [9] for the topic considered in § 3.

2. Quadrangle inequalities.

Example 1. Given a set of points X on the plane, how do we find five points that span a pentagon with maximum perimeter?

A natural solution based on dynamic programming would be to seek out maximum triangles, maximum quadrilaterals, and maximum pentagons in turn. It is not difficult to argue that we can restrict our consideration to the extreme points of X. Therefore let us assume the convex hull of X to be $P = \langle v_1, v_2, \cdots, v_n \rangle$, and the distance between v_i and v_j to be d_{ij}. Then maximum triangles can be found by computing the largest entry in the matrix $D + D \otimes D$, where $D = (d_{ij})$, and $\otimes$ denotes the (max, +)-multiplication of two matrices defined by

$$F \otimes G = (p_{ij}), \quad \text{where } p_{ij} = \max \{f_{ik} + g_{kj} | i \leqq k \leqq j\} \text{ for } F = (f_{ij}) \text{ and } G = (g_{ij}).$$

Since $\otimes$ is associative, we will write D^2 for $D \otimes D$, and D^t for $D^{t-1} \otimes D$. A maximum pentagon then corresponds to a maximum entry in $D + D^4$, where D^4 may be evaluated as $D^2 \otimes D^2$. In general, a maximum t-gon can be found by first computing D^{t-1} and then finding a maximum entry in $D + D^{t-1}$. Since D^{t-1} can be obtained from D in $O(\log t)$(max, +)-multiplications (see [7, § 4.6.3], for example) at a cost of $O(n^3)$ steps per matrix multiplication, the answer can be obtained in $O(n^3 \log t)$ steps.

Now we pose the question: Can $D \otimes D$ be computed in time faster than $O(n^3)$? It turns out that, by properties of the Euclidean metric d_{ij}, if we let $K(i, j)$ denote

* Received by the editors February 25, 1982.
† Xerox Palo Alto Research Center, Palo Alto, California 94304.

$\max\{k|d_{ik}+d_{kj}=(D\otimes D_{ij}\}$, then $K(i,j)$ is a monotone function of i and j (see Theorem 1).

CLAIM 1. $K(i,j)\leqq K(i,j+1)\leqq K(i+1,j+1)$.

This property enables us to limit our search for the optimal k, while computing $(D\otimes D)_{i,j+1}=\max\{d_{i,k}+d_{k,j+1}|i\leqq k\leqq j+1\}$, to those k that lie between $K(i,j)$ and $K(i+1,j+1)$, provided that the latter two values are already known. This suggests computing $(D\otimes D)_{ij}$ by diagonals, in order of increasing values of $j-i$. The cost for computing all entries of one diagonal is $O(n)$ as a result of Claim 1 and the total cost for obtaining $D\otimes D$ is thus only $O(n^2)$.

More generally, when one forms the product $D^r\otimes D^s$ for any $r\geqq 1$ and $s\geqq 1$, monotonicity properties analogous to Claim 1 also hold (Theorems 1 and 2). This implies that our earlier dynamic programming algorithm for finding maximum t-gons can be speeded up from $O(n^3\log t)$ to $O(n^2\log t)$.

The critical property of the Euclidean distance function d_{ij} that makes Claim 1 true is what we call the "quadrangle inequalities". We say that a real-valued function $f(i,j)$, where $1\leqq i\leqq j\leqq n$, satisfies *convex quadrangle inequalities* (*convex* QI) if

$$f(i,k)+f(j,l)\geqq f(i,l)+f(j,k)\quad \text{for } i\leqq j\leqq k\leqq l.$$

The same inequalities with signs reversed are called *concave quadrangle inequalities* (*concave* QI):

$$f(i,k)+f(j,l)\leqq f(i,l)+f(j,k)\quad \text{for } i\leqq j\leqq k\leqq l.$$

Example 2. It is easy to see that the distance function d_{ij} for vertices of a convex polygon in Example 1 satisfies the convex QI. Some other examples of functions are given below.

$f(i,j)=a_i+a_{i+1}+\cdots+a_j$
$f(i,j)=a_i+a_{i+1}+\cdots+a_{j-1}$
$f(i,j)=a_{i+1}+a_{i+2}+\cdots+a_{j-1}$
} all satisfy both concave QI and convex QI;

$f(i,j)=a_i\cdot a_{i+1}\cdot\cdots\cdot a_j$ satisfies concave QI if all a_k's are $\geqq 1$.

Furthermore, convex QI are preserved by convex, nondecreasing mappings (for example, $\log d_{ij}$ satisfies convex QI); while concave QI are preserved by concave, nondecreasing mappings (for example, $f^2(i,j)$ satisfies concave QI for any of the four f's defined above). Additional QI-preserving mappings that are of particular importance to dynamic programming will be discussed in Theorem 2 and 3 below.

Our earlier Claim is derived from the following general theorem. Let $K_{f\otimes g}(i,j)$ denote $\max\{k|f(i,k)+g(k,j)=(f\otimes g)(i,j)\}$; that is, $K_{f\otimes g}(i,j)$ is the largest index k for which $f(i,k)+g(k,j)$ achieves the maximum. For simplicity, we will write $K(i,j)$ for $K_{f\otimes g}(i,j)$ whenever the context $f\otimes g$ is understood.

THEOREM 1. *If both f and g satisfy convex* QI, *then $K_{f\otimes g}(i,j)$ is a monotone function of i and j*:

$$K(i,j)\leqq K(i+1,j)\leqq K(i+1,j+1).$$

As we saw in Example 1, the above theorem allows us to compute $K_{f\otimes g}$ and $f\otimes g$ with a cost of only $O(n)$ per diagonal, thus $O(n^2)$ in total.

COROLLARY A. *If both f and g satisfy convex* QI, *then $f\otimes g$ and $K_{f\otimes g}$ can be computed in $O(n^2)$ time and space.*

The above results regarding convex QI and maximization problems have parallels in concave QI and minimization problems. Define

$$f \oslash g(i,j) = \min\{f(i,k)+g(k,j) | i \leqq j \leqq k\}.$$

COROLLARY B. *If both f and g satisfy concave* QI, *then* $f \oslash g$ *and* $K_{f \oslash g}$ *can be computed in* $O(n^2)$ *time and space.*

The following theorem allows us to apply these corollaries iteratively, in situations such as Example 1.

THEOREM 2. *If both f and g satisfy convex* QI, *then* $f \otimes g$ *also satisfies convex* QI. *If both f and g satisfy concave* QI, *then* $f \oslash g$ *also satisfies concave* QI.

We also find the concepts of QI useful in the evaluation of recurrence relations involving either minimization or maximization operations. We will mention one such result for concave QI.

A function $w(i,j)$ where $i \leqq j$ is said to be *monotone* if it is monotonically increasing on the lattice of intervals (ordered by inclusion), i.e.,

$$w(i,j) \leqq w(i',j') \quad \text{if } [i,j] \subseteq [i',j'].$$

THEOREM 3. *Let* $c(i,j)$, *where* $i \leqq j$, *be defined by*

$$c(i,j) = w(i,j) + \min_{i<j\leqq k} [c(i,k-1)+c(k,j)] \quad \text{if } i<j,$$
$$c(i,i) = a(i). \tag{1}$$

If w satisfies concave QI *and is monotone, then c satisfies concave* QI.

In consequence, we have the following speed-up result analogous to Theorem 1 and its corollaries.

COROLLARY. *For a function* $c(i,j)$ *satisfying the description of Theorem* 3, *we can compute* $K_{c \oslash c}(i,j)$ *and* $c(i,j)$ *for* $0 \leqq i \leqq j \leqq n$ *in* $O(n^2)$ *time and space.*

Example 3. A bookstore is interested in organizing its index files in a way to facilitate look-ups. Take the subject index for example. Suppose that the index, alphabetically ordered, consists of a number of key subjects such as {ART, COOKING, $\cdots$, TRAVEL}, plus other subjects that fall in the intervals in between, namely {A-ART, ART-COOKING, $\cdots$}. We will denote the key subjects by $\{K_1, K_2, \cdots, K_n\}$, and the intervals by $\{I_0, I_1, \cdots, I_n\}$. Assume that the access probability for key K_i is p_i, and that for interval I_j is q_j. We would like to build a binary tree, with the K_i's as internal nodes, and the I_j's as external nodes, such that the expected

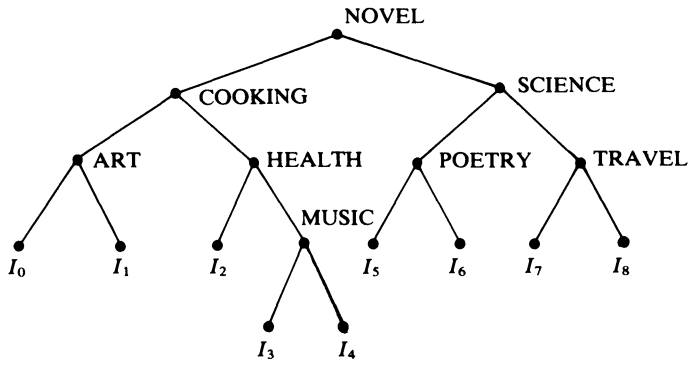


FIG. 1. *A binary search tree.*

number of comparisons in looking up a subject, namely

$$\sum_{1\leqq i\leqq n} p_i(1+\text{level of } K_i)+\sum_{0\leqq j\leqq n} q_j(\text{level of } I_j)$$

is minimized (Fig. 1).

Since all subtrees of an optimal tree must themselves be optimal, this problem can be solved by dynamic programming. One naturally arrives at recurrence relations of the form (1), with $c(i, j)$ being the minimum cost of a subtree for keys $\{K_{i+1}, \cdots, K_j\}$ and intervals $\{I_i, \cdots, I_j\}$, and

$$\begin{aligned} &w(i, j)=p_{i+1}+\cdots+p_j+q_i+\cdots+q_j, \\ &a(i)=0. \end{aligned} \tag{2}$$

The cost of the optimal tree that we are interested in is $c(0, n)$. As noted in Example 2, the function $w(i, j)$ in (2) satisfies concave QI. Therefore by the corollary to Theorem 3, we can compute the values of $c(i, j)$ in $O(n^2)$ time and space. Furthermore, once $c(0, n)$and $K_{c\oslash c}(0, n)$ are found, we can then trace the information in $K_{c\oslash c}(i, j)$ "from top down" to obtain the actual construction of an optimal binary tree in $O(n)$ steps.

Remarks. The problem of optimal binary search trees discussed above is a classical example of dynamic programming in the computer science literature. The original $O(n^3)$ solution by setting up the recurrence relations (1) was due to Gilbert and Moore [3]. Then Knuth [5] showed that the algorithm can be speeded up to $O(n^2)$ by proving that $K_{c\oslash c}(i, j)$ is monotone. However, his proof of monotonicity was given for the particular $w(i, j)$ as defined by (2), and thus not apparently generalizable. For the problem considered in Example 1, some recent results can be found in [2].

3. Multiplying a sequence of matrices. We now turn to another example of a classical dynamic programming algorithm [1] which saw much notable progress lately.

Example 4. Let $M_1, M_2, \cdots, M_n$ be n matrices of dimensions $d_1\times d_2$, $d_2\times d_3, \cdots, d_n\times d_{n+1}$, respectively. What is the optimal order, by multiplying two matrices at a time, for evaluating the product $M_1\times M_2\times\cdots\times M_n$?

To be more specific, let us assume that the cost for multiplying a $p\times q$ matrix with a $q\times r$ matrix is pqr. Consider, for example, four matrices $M_1, \cdots, M_4$ of dimensions 100×1, 1×50, 50×20 and 20×1. Evaluating their product in the left-to-right order $((M_1\times M_2)\times M_3)\times M_4$ would cost 125,000 operations, while the minimum cost, achieved by $M_1\times((M_2\times M_3)\times M_4)$, is only 2,200.

Using dynamic programming, a solution to this problem can be obtained by defining $c(i, j)$ to be the minimum cost for evaluating $M_i\times M_{i+1}\times\cdots\times M_j$, and setting up the recurrence relations

$$c(i, j)=\min_{i<k\leqq j}[c(i, k-1)+c(k, j)+d_id_kd_j] \quad \text{if } i<j,$$

$$c(i, i)=O.$$

This gives an $O(n^3)$ algorithm for computing $c(1, n)$. Can we do any better? As our tools based on quadrangle inequalities (Theorem 3) do not apply to recurrence relations of the present form, we must come up with a different technique.

In the following, we will first develop a geometric representation for the problem. Then, by looking at the case $n=3$, we will extrapolate some simple properties of the optimal solution. We then show how these properties can be utilized to lead to an $O(n^2)$ dynamic programming algorithm.

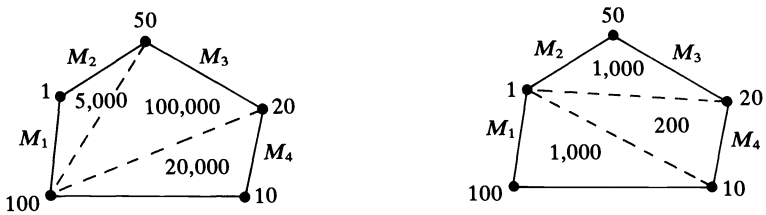


FIG. 2. *Geometric representation for the evaluation of a matrix chain.*

We will use the vertices of an $(n+1)$-sided convex polygon in the plane to represent the $n+1$ parameters $\langle d_1, d_2, \cdots, d_{n+1}\rangle$. A directed edge from d_i to d_j, where $i<j$, will be interpreted as a matrix of dimension $d_i \times d_j$, representing the product $M_i \times M_{i+1} \times \cdots \times M_{j-1}$. Thus, the $n+1$ sides of the convex polygon correspond to the n input matrices and the final product, while any chord represents a potential partial product. It is easy to see that there is a one-to-one correspondence between the different ways of parenthesizing $M_1 \times M_2 \times \cdots \times M_n$ and the possible ways of triangulating the polygon $\langle d_1, d_2, \cdots, d_{n+1}\rangle$. If we associate a cost of $d_i d_k d_j$ with a triangle whose vertices are labeled d_i, d_k and d_j, then our original problem becomes the problem of finding an optimal triangulation of the polygon $\langle d_1, d_2, \cdots, d_{n+1}\rangle$. Figure 2 illustrates the triangulations corresponding to the two different ways of evaluating $M_1 \times M_2 \times M_3 \times M_4$ mentioned earlier.

From now on, we will refer to the d_i's as *weights*. Let $w_1 \leqq w_2 \leqq \cdots \leqq w_n$ be the weights of an n-sided convex polygon P sorted into nondecreasing order. (The ordering may not be unique as some of the weights may be equal; we assume that a particular ordering is chosen and remains fixed.) We will use $w_i w_j$ to denote a directed edge from w_i to w_j, and $w_i w_j w_k$ to denote a triangle with vertices w_i, w_j and w_k, when there is no ambiguity to these notations. We will also use the term partition interchangeably with triangulation.

Consider the case of a quadrilateral. If w_1 and w_2 face each other, then the arc $w_1 w_2$ gives us an optimal partition. This is so because

$$w_1 \cdot w_2 \cdot w_4 + w_1 \cdot w_2 \cdot w_3 \leqq w_2 \cdot w_3 \cdot w_4 + w_1 \cdot w_3 \cdot w_4,$$

or

$$1/w_3 + 1/w_4 \leqq 1/w_1 + 1/w_2.$$

Similarly, if w_1 faces w_3, then $w_1 w_3$ is an optimal partition, because

$$1/w_2 + 1/w_4 \leqq 1/w_1 + 1/w_3.$$

On the other hand, if w_1 faces w_4, then either $w_1 w_4$ or $w_2 w_3$ could be optimal.

The above generalizes to an n-gon by an inductive argument.

LEMMA 1. *Let P be an n-gon with weights $w_1 \leqq w_2 \leqq \cdots \leqq w_n$. Then there exists an optimal partition π for which the following is true.*

(a) *w_1 and w_2 are adjacent (either by a side edge or by a chord); similarly for w_1 and w_3.*

(b) *if both $w_1 w_2$ and $w_1 w_3$ are side edges, then either $w_1 w_4$ or $w_2 w_3$ exists as a chord.*

Lemma 1 implies that we can set up the following recursive procedure for finding an optimal triangulation. We use P_{ij} to denote the subpolygon of P consisting of those vertices lying between w_i and w_j in a clockwise traversal.

```
    PROCEDURE Partition [P]
begin
    if |P| = 1 or 2 then return ∅
    else
        if P is a triangle then return P
    else
        if w1 and w2 are not adjacent then return Partition [P1,2] ∪ Partition [P2,1]
    else
        if w1 and w3 are not adjacent then return Partition [P1,3] ∪ Partition [P3,1]
    else
        return better of {Partition [P2,3] ∪ Partition [P3,2],
                          Partition [P1,4] ∪ Partition [P4,1]};
    end.
```

As it is, this recursive algorithm requires exponential time, since in the worst case the last **else** clause could generate two problems of size $n-c$ for some constant c. We will show that, however, the total number of calls on *distinct* subpolygons $\{Q\}$ is bounded by $O(n^2)$. Furthermore, these $O(n^2)$ subpolygons can be ordered in such a way that in computing Partition[Q], solutions to its subproblems are already available. In other words, one can turn Partition into a dynamic programming algorithm with an $O(n^2)$ space and time bound. To this end, we need a characterization of those chords w_iw_j in the original polygon P that may arise as w_2w_3 in some recursive call spawned by Partition [P].

DEFINITION. A (directed) chord w_iw_j of P is called a *bridge*, if all weights w_k in P_{ij} satisfy $k \geqq \max\{i, j\}$.

Note that both w_1w_2 and w_2w_1 are bridges, and it is the only instance where two bridges correspond to the same (undirected) edge. The side edges of P may be viewed as degenerate bridges, henceforth we will include them in the definition for convenience.

It is easy to check that bridges have the following properties:

1. Two bridges never intersect (except possibly at the endpoints); therefore there are at most $O(n)$ bridges.

2. A partial order $\prec$ can be imposed on the set of bridges if we define $w_{i'}w_{j'} \prec w_iw_j$ to mean $P_{i'j'} \subseteq P_{ij}$.

3. The transitive reduction of $\prec$ (i.e., the subgraph of $\prec$ with all edges implied by transitivity removed) is a forest, for $a \prec b$ and $a \prec c$ imply that b and c are comparable in $\prec$. We shall denote this forest by $T[\prec]$. Note that w_1w_2 and w_2w_1 are the two roots of $T[\prec]$, and the leaves are the degenerate bridges (sides) of P.

4. Any nonleaf node w_iw_j of $T[\prec]$ has exactly two sons, namely w_iw_k and w_kw_j where k is the smallest index (aside from i and j) in P_{ij}; we will refer to them respectively, assuming $i<j$, as the *minson* and the *maxson* of w_iw_j. Thus $T[\prec]$ is actually the union of two binary trees. Figure 3 gives an example of a polygon P and the corresponding $T[\prec]$.

Procedure MarkBridges below identifies and outputs the bridges of P as it makes one clockwise scan of the weights. The bridges are actually generated in (slightly modified) postorder [6] of the tree $T[\prec]$; therefore, in particular, they are topologically sorted into a nondecreasing order consistent with $\prec$. The procedure employs a stack S, and we use the notations $S \Leftarrow x$ abd $x \Leftarrow S$ for pushing and popping as defined in [6].

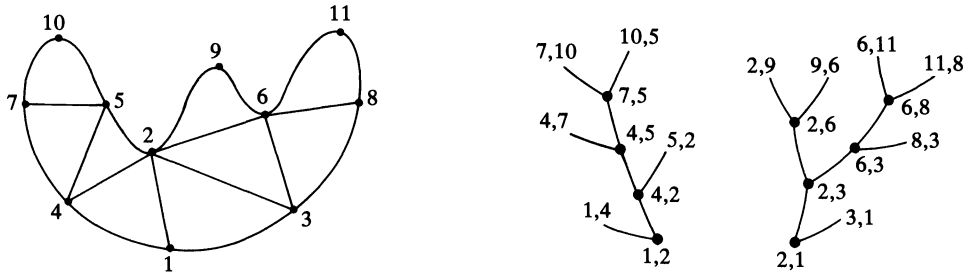


FIG. 3. *A polygon P and the corresponding forest T[<]. (The weights are represented by their indices only.)*

```
PROCEDURE MarkBridges [P];
begin
    find the minimum weight w_1;
    w ← w_1;
    repeat
        begin
            S ⇐ w;
            w ← nextweight;                    —Going clockwise from w_1.
            while top(S) > w do
                begin t ⇐ S;
                    output (top(S), t) and (t, w) as bridges;
                end;
        end
    until w = w_1;                             —Halt after returning to w_1.
end.
```

DEFINITION. A subpolygon Q of P is called a *cone*, if $Q = P_{ij} \cup w_i w_j w_k$ where $b = w_i w_j$ is a bridge of P, and $k \leqq \min\{i, j\}$. We also denote a cone Q by (b, w_k) (Fig. 4).

In particular, P_{ij} for any bridge $w_i w_j$ is a cone, and P itself is the union of two cones $P_{1,2}$ and $P_{2,1}$. The existing partial orders on bridges and on weights induce a natural alphabetic order on cones.

DEFINITION. We say that a cone $Q' = (b', w_{k'})$ *precedes* a cone $Q = (b, w_k)$ if either (1) $b' \prec b$, or (2) $b' = b$ and $k' \geqq k$.

LEMMA 2. *Any subpolygon that may arise in the execution of* Partition $[Q]$, *for a cone* $Q = (b, w_k)$, *is either a triangle or a cone* Q' *which precedes* Q.

Thus we can use dynamic programming to compute and tabulate the solutions to all cones in accordance with their precedence order. The actual recurrence relations

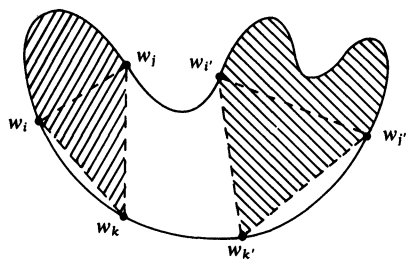


FIG. 4. *Example of two cones (the shaded regions).*

have been incorporated into the following program. We use Partition $[Q]$ to refer to the table entry containing the optimal solution to cone Q. The outer **for** loop iterates over all b in the order as they are generated by MarkBridges, while the inner **for** loops iterate over all w_k with $k \leqq i$ in decreasing order. The algorithm runs in $O(n^2)$ time, as there are at most $2n$ bridges, and at most n cones for a given bridge.

```
PROCEDURE DP-Partition [P]
begin
    for b = w_i w_j ∈ B do          —B is the output of Markbridges [P].
    begin                           —Assume that i < j.
        if b is a leaf then
            for all cones Q = (b, w_k⟩ with k ≦ i do
                if w_k = w_i then Partition [Q] ← ∅          —Q = w_i w_j
                             else Partition [Q] ← Q;         —Q = w_i w_j w_k
        if b is not a leaf then
            for all cones Q = (b, w_k⟩ with k ≦ i do
                if w_k = w_i then Partition [Q] ← Partition[(minson (b), w_i⟩] ∪
                                                  Partition[(maxson (b), w_i⟩]
                             else Partition [Q]
                             ← better of {Partition [(b, w_i⟩] ∪ w_i w_j w_k,
                                          Partition[(minson (b), w_k⟩] ∪
                                          Partition [(maxson (b), w_k⟩]};
    end;
    Partition [P] ← Partition [P_{1,2}] ∪ Partition [P_{2,1}];
end.
```

Remark: In 1980, Hu and Shing [4] gave an $O(n \log n)$ algorithm for solving this problem. However, their presentation is exceedingly long; a more concise exposition, including the preceding algorithm, can be found in Yao [9].

4. Conclusions. We surveyed some recent results in the design of dynamic programming algorithms. These results illustrate two approaches for obtaining speed-up in dynamic programming: one general and the other problem specific. In the first case, the quadrangle inequalities provide a type of sufficient conditions by which speed-up is guaranteed, and these conditions apply to a broad class of problems. In the second case, we present a nonobvious algorithm for solving the matrix chain product problem efficiently. Even though the techniques involved in the second case are problem specific, it serves as an excellent example for illustrating how speed-up comes about in dynamic programming: namely, by trying to solve individual subproblems fast, and by trying to keep small the total number of distinct subproblems that need solving.

REFERENCES

[1] A. AHO, J. HOPCROFT AND J. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading MA, 1974.

[2] J. E. BOYCE, D. P. DOBKIN, R. L. DRYSDALE, III AND L. J. GUIBAS, *Finding extremal polygons*, Proc. 14th Annual ACM Symposium on Theory of Computing, (1982), pp. 282–289.

[3] E. N. GILBERT AND E. F. MOORE, *Variable length encodings*, Bell System Tech. J., 38 (1959), pp. 933–968.

[4] T. C. HU AND M. T. SHING, *Computation of matrix chain products, Part I and II*, manuscript (1981). (Extended abstract in Proc. 21st Annual Symposium on Foundations of Computer Science, 1980, pp. 28–35.)

[5] D. E. KNUTH, *Optimum binary search trees*, Acta Informatica, 1 (1971), pp. 14–25.
[6] ——, *The Art of Computer Programming, Vol 1: Fundamental Algorithms*, second ed., Addison-Wesley, Reading, MA, 1975.
[7] ——, *The Art of Computer Programming, Vol 2: Seminumerical Algorithms*, second ed., Addison-Wesley, Reading, MA, 1981.
[8] F. F. YAO, *Efficient dynamic programming using quadrangle inequalities*, Proc. 12th Annual ACM Symposium on Theory of Computing, (1980), pp. 429–435.
[9] ———, *A note on optimally multiplying a sequence of matrices*, manuscript (1982).

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0016 $01.00/0

AN ALGORITHM FOR PARTITIONING THE NODES OF A GRAPH*

EARL R. BARNES†

Abstract. Let $G=\{N, E\}$ be an undirected graph having nodes N and edges E. We consider the problem of partitioning N into k disjoint subsets $N_1, \cdots, N_k$ of given sizes $m_1, \cdots, m_k$, respectively, in such a way that the number of edges in E that connect different subsets is minimal. We obtain a heuristic solution from the solution of a linear programming transportation problem.

1. Introduction. The problem of partitioning the nodes of a graph arises in the laying out of circuits on computer boards [1, Chap. 7], [2], [3], computer program segmentation [4], [5], [6, pp. 74–126], and in several other areas. In each case an undirected graph having n nodes $N=\{1, 2, \cdots, n\}$ and $|E|$ edges E is given. Also given are k positive integers $m_1 \geqq m_2 \geqq \cdots \geqq m_k$ satisfying $\sum_{i=1}^{k} m_i = n$. The problem is to partition the nodes N into k disjoint subsets $N_1, \cdots, N_k$ of sizes $m_1, \cdots, m_k$, respectively, in such a way that the number of edges connecting different subsets is minimal. An edge which connects two distinct subsets is said to be cut by the partition.

In this paper we show that the partitioning problem is equvalent to a matrix approximation problem. We show that an approximate solution of this matrix approximation problem can be obtained by solving a linear programming problem. The solution of the linear programming problem gives a partition of the nodes which, at least heuristically, cuts a number of edges close to the minimim. A good survey of previous approaches to the partitioning problem is given in [7].

2. The algorithm. Let a_{ij} be the number of edges connecting nodes i and j, $i \neq j$, and let $a_{ii}=0$, $i=1, \cdots, n$. Let A denote the $n \times n$ matrix (a_{ij}). A is the adjacency matrix of the graph. Given a partition of the graph, let $P=(p_{ij})$ be the $n \times n$ matrix defined by

$$p_{ij} = \begin{cases} 1 & \text{if nodes } i \text{ and } j \text{ belong to the same subset,} \\ 0 & \text{otherwise.} \end{cases}$$

In this way, we identify each partition with a matrix P.

Let P be any partition and let E_{nc} denote the number of edges not cut by P. let E_c denote the number of edges cut. Then clearly

$$2E_{nc} = \sum_{i \neq j} a_{ij} p_{ij} = \sum a_{ij} p_{ij},$$

and

$$2E_c = 2(|E| - E_{nc}) = 2|E| - \sum a_{ij} p_{ij}.$$

Let $\|C\| = \sqrt{\sum |c_{ij}|^2}$ denote the Frobenius norm of a matrix $C=(c_{ij})$. Each partition P has m_i rows containing exactly m_i ones, $i=1, \cdots, k$. Thus $\|P\|^2 = \sum m_i^2$ for each partition P.

Observe that

$$\|A-P\|^2 = \|A\|^2 - 2\sum a_{ij} p_{ij} + \|P\|^2 = 4E_c + \|A\|^2 + \sum m_i^2 - 4|E|.$$

This shows that the partition which minimizes the number of edges cut is the one nearest A. Thus our original problem has been reduced to approximating A by a partition. A useful estimate of how close a partition can be to A is provided by the

* Received by the editors March 11, 1981, and in revised form April 30, 1982.

† IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.

Hoffman–Wielandt inequality [8]. According to this inequality if A and B are real $n \times n$ symmetric matrices with eigenvalues $\alpha_1 \geqq \cdots \geqq \alpha_n$ and $\beta_1 \geqq \cdots \geqq \beta_n$ respectively, then

$$\|A-B\|^2 \geqq \sum_{i=1}^{n} (\alpha_i - \beta_i)^2.$$

Let $\lambda_1 \geqq \cdots \geqq \lambda_n$ denote the eigenvalues of the adjacency matrix A. The eigenvalues of each partition P are given by $m_1, \cdots, m_k, 0, \cdots, 0$. To see this, observe that the rows (columns) of P corresponding to nodes in the same subset N_j are identical. Thus each partition has exactly k distinct rows (columns). Moreover, the distinct rows of P are mutually orthogonal. This follows from the disjointness of the sets $N_1, \cdots, N_k$. It follows that P has rank k. Thus 0 is an $(n-k)$-fold eigenvalue of each partition. *Since the distinct rows (columns) of each partition are mutually orthogonal* $0-1$ *vectors they are eigenvectors of the partition.* They correspond to the eigenvalues $m_1, \cdots, m_k$. We are now in a position to apply the Hoffman–Wielandt inequality. If A is the adjacency matrix of the graph, this inequality states that

$$\|A-P\|^2 \geqq \sum_{i=1}^{k} (\lambda_i - m_i)^2 + \sum_{i=k+1}^{n} \lambda_i^2 \tag{2.1}$$

for any partition P.

The set of orthonormal eigenvectors of a partition P obtained from the k distinct columns of P can be written as

$$\nu_j = \pm \frac{1}{\sqrt{m_j}} \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}, \qquad j = 1, \cdots, k, \tag{2.2}$$

where each x_{ij} is either 0 or 1 and

$$\sum_{i=1}^{n} x_{ij} = m_j, \quad j = 1, \cdots, k, \qquad \sum_{j=1}^{k} x_{ij} = 1, \quad i = 1, \cdots, n. \tag{2.3}$$

Let $u_1, \cdots, u_n$ denote a set of orthonormal eigenvectors of A corresponding to the eigenvalues $\lambda_1, \cdots, \lambda_n$, respectively. We say that $u_1, \cdots, u_k$ are the first k eigenvetors of A. Let U denote the $n \times n$ matrix whose jth column is u_j and let $\Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$. Let V denote the $n \times k$ matrix whose jth column is ν_j given by (2.2). Assume for the moment that the + sign in (2.2) is chosen. The correct choice of signs will be discussed later. Let $M = \operatorname{diag}(m_1, \cdots, m_k)$. We can write

$$A = U\Lambda U^T \quad \text{and} \quad P = VMV^T.$$

This gives

$$\|A-P\|^2 = \|U\Lambda U^T - VMV^T\|^2 = \|\Lambda - U^T VM(U^T V)^T\|^2.$$

If we could choose V such that

$$U^T V = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ & & \ddots & \vdots \\ \vdots & & & 1 \\ 0 & & \cdots & 0 \\ & & \cdots & \\ 0 & & \cdots & 0 \end{pmatrix} \triangleq J, \tag{2.4}$$

we would have equality in (2.1), and P would be the best approximation of A by a partition. However, it is generally impossible to choose V such that (2.4) is satisfied, due to the constraints (2.3). We therefore choose V to minimize the error $\|U^TV-J\|$ subject to (2.3). We have

$$(2.5)\quad \|U^TV-J\|^2=\|V-UJ\|^2=\sum_{j=1}^{k}\|v_j-u_j\|^2=\sum_{j=1}^{k}\left(2-2\sum_{i=1}^{n}v_{ij}u_{ij}\right)=2k-2\sum_{i=1}^{n}\sum_{j=1}^{k}v_{ij}u_{ij}$$

where u_{ij} and v_{ij} are the ith components of the vectors u_j and v_j respectively. Substituting for v_j from (2.2), we see that the V which minimizes (2.5) is obtained by solving the transportation problem

$$(2.6)\qquad \text{minimize}-\sum_{i=1}^{n}\sum_{j=1}^{k}\left(\frac{u_{ij}}{\sqrt{m_j}}\right)x_{ij}$$

subject to

$$\sum_{i=1}^{n}x_{ij}=m_j,\qquad j=1,\cdots,k,$$

$$\sum_{j=1}^{k}x_{ij}=1,\qquad i=1,\cdots,n,$$

$$x_{ij}\geqq 0,\qquad i=1,\cdots,n,\quad j=1,\cdots,k.$$

Since the numbers $m_1,\cdots,m_k,1$, are integers, a vertex solution of this problem will be integer valued. And since $\sum_{j=1}^{k}x_{ij}=1$, each x_{ij} will be 0 or 1. Thus, given a solution of the linear programming problem (2.6), we obtain all the distinct columns (2.2) of a partition P. In terms of the nodes the partition is given by $N=\bigcup_{j=1}^{k}N_j$ where

$$(2.7)\qquad N_j=\{i|x_{ij}=1\},\qquad j=1,\cdots,k.$$

An algorithm for solving (2.6) is described in [9] and in most books on linear programming. To complete our description of the partitioning algorithm, we must explain how the signs in (2.2) are chosen. Clearly the most desirable choice of signs is the one for which the corresponding transportation problem has the smallest possible minimum. One way of determining this choice is to solve the 2^k transportation problems corresponding to all possible choices of signs in (2.2) and select the one with the smallest minimum value. A less tedious, but heuristic, procedure for choosing the correct signs will now be described.

As far as the linear programming problem (2.6) is concerned, changing the sign of v_j is equivalent to changing the signs of the components $u_{1j},\cdots,u_{nj}$ of u_j in (2.6). Thus we can fix the signs of the v_j's and concentrate on changing the signs of the u_j's so as to make the minimum (2.6) as small as possible. It is convenient to choose the + sign for each v_j in (2.2). The v_j's then lie in the nonnegative orthant of n-dimensional Euclidean space. Since the v_j's are to be chosen to approximate the u_j's, we choose the u_j's to be as close to the nonnegative orthant as possible. Thus given a set of orthonormal first-k eigenvectors $u_1,\cdots,u_k$ of A, we obtain an equivalent set by replacing u_j by $-u_j$ whenever $-u_j$ has a larger projection on the nonnegative orthant than u_j. We use this new set of eigenvectors of A to determine a partition (2.7) of N as described above. The projection of a vector $y=(y_1,\cdots,y_n)$ on the nonnegative orthant is given by $y^+=(y_1^+,\cdots,y_n^+)$ where $y_i^+=\max\{y_i,0\}$. The size of the projection is measured by $\|y^+\|=\sqrt{\sum(y_i^+)^2}$.

3. An alternative partitioning procedure. The discussion of this section applies only to the case $k \geqq 3$. Suppose we have found a partition $N = \bigcup_{j=1}^{k} N_j$ of the nodes of a graph by the procedure described in the previous section. If for each j, ν_j is a good approximation of u_j, then it is clear tht the partition we have found is very close to an optimal one. However, it may happen that for some j, ν_j is a not a good approximation of u_j. In this case the partition we have found may differ significantly from an optimal one. Typically ν_j will fail to be a good approximation of u_j when u_j has significantly more than m_j components greater than $\frac{1}{2}\sqrt{1/m_j}$. In such a case, there is really no good approximation of u_j by a vector of the form (2.2). The number of positive components of such a ν_j is too small for ν_j to be a good approximation of u_j. One way to remedy this situation is to change the problem so that ν_j has more nonzero components. If $j = 1$ or if $m_j + m_k \leqq m_{j-1}$, this can be accomplished by using our procedure to find an $m_1, \cdots, m_{j-1}, m_j + m_k, \cdots, m_{k-1}$ partition. The ν_j for this problem will have $m_j + m_k$ nonzero components. It is likely to give a better approximation of u_j than the original ν_j. To obtain an $m_1, \cdots, m_k$ partition we simply find an m_j, m_k partition of the set N_j obtained in the $m_1, \cdots, m_{j-1}, m_j + m_k, \cdots, m_{k-1}$ partition. In some cases this partition will be superior to the original $m_1, \cdots, m_k$ partition. We demonstrate this with an example in the next section. Observe that the m_j, m_k partition is on a graph containing only $m_j + m_k$ nodes. Thus the ν's that appear have a larger percentage of nonzero components than the ν's that appeared in the original problem.

In some cases it may be desirable to increase the number of nonzero components in several of the ν_j's. This can be done by a natural extension of the procedure just described. Thus, for example, an m_1, m_2, m_3, m_4, m_5 partition may be found by first finding an $m_1 + m_4, m_2 + m_5, m_3$ partition and then m_1, m_4 and m_2, m_5 partitions.

We close this section by giving an explicit condition for when the approximation of u_j by ν_j can be improved if m_j is replaced by $m_j + m_k$ and m_k by 0 in the original problem. The approximation can be improved if

$$\sum_{i \in N-(N_j \cup N_k)} u_{ij}^2 + \sum_{i \in N_j \cup N_k} (1/\sqrt{m_j + m_k} - u_{ij})^2 < \sum_{i \in N - N_j} u_{ij}^2 + \sum_{i \in N_j} (1/\sqrt{m_j} - u_{ij})^2,$$

or, equivalently, if

$$\sum_{i \in N_j \cup N_k} u_{ij}/\sqrt{m_j + m_k} > \sum_{i \in N_j} u_{ij}/\sqrt{m_j}. \tag{3.1}$$

This inequality exhibits an approximation of u_j in the new problem which is an improvement of the old approximation.

To implement our method we need an algorithm for computing the first k eigenvectors of a real symmetric matrix A. Such an algorithm is described in [10]. It is the block Lanczos method.

4. Obtaining a local minimum for E_c. In general a partition obtained by the methods of the previous two sections will not be optimal, even locally. That is, given such a partition, it may be possible to decrease E_c by interchanging a node in one of the sets N_j with a node in another of these sets. The possibility of improving a partition by such interchanges should always be investigated when our method is applied. Interchanges of single nodes should be carried out as long as it is possible to do so without increasing E_c. However, in making interchanges that leave E_c fixed, care must be taken to avoid cycling, that is, returning to a partition that has already been examined. When it is no longer possible to make a single node interchange without increasing E_c, or cycling, a local minimum for E_c has been obtained. In general, E_c will have many local minima. However, we expect the one found by the procedure

we just described to be very close to a global minimum. This is because we expect our original partition to be close to an optimal one.

Given an initial partition $N = \bigcup_{j=1}^{k} N_j$, a simple formula for computing the change in E_c due to a single node interchange between two of these subsets is given in [7, p. 296]. We repeat it here for the sake of the reader. Suppose we wish to interchange a node in N_1 with one in N_2. For each $i \in N_1$ let $E_1(i)$ denote the number of edges connecting i to nodes in N_2. That is, $E_1(i) = \sum_{j \in N_2} a_{ij}$. Let $I_1(i) = \sum_{j \in N_1} a_{ij}$ denote the number of edges connecting i to nodes in N_1. Let $D_1(i) = E_1(i) - I_1(i)$. $D_1(i)$ is the amount by which E_c is reduced if node i is switched from N_1 to N_2. Define $E_2(j)$, $I_2(j)$, and $D_2(j)$ similarly for each node $j \in N_2$. The amount by which E_c is reduced if node $i \in N_1$ is interchanged with $j \in N_2$ is

$$D(i, j) = D_1(i) + D_2(j) - 2a_{ij}.$$

If $D(i, j)$ is negative for each $i \in N_1$ and each $j \in N_2$, then it is not possible to make a single node interchange without increasing E_c. If some $D(i, j)$ is $\geqq 0$ the nodes i and j for which $D(i, j)$ is a maximum should be interchanged provided this maximum is positive. If the maximum is zero, an interchange should be made only if it avoids cycling.

5. Examples. To demonstrate our procedure we partition two graphs that appear in [11, p. 425]. The first graph has 20 nodes whose connections are described in Table 1. First we partition the nodes into two sets containing 10 nodes each. The first step of our procedure requires that we find the first two eigenvectors of the adjacency matrix A. Let the first two eigenvectors of A, normalized to have unit length, be denoted by u_1 and u_2 respectively. Orient these vectors so that their projections on the nonnegative orthant are maximal. Form the transportation tableau whose first and second rows contain the vectors $-\sqrt{1/10}u_1$ and $-\sqrt{1/10}u_2$ respectively. This tableau is shown in Table 2. We have rounded all numbers to three decimal places. The rows in the tableau correspond to two origins, each having 10 units of a product for shipment. The columns correspond to 20 destinations, each requiring 1 unit of the product. The numbers $-\sqrt{1/10}u_{ij}$ appearing in the upper portion of the squares in the tableau represent the cost of shipping a unit of the product from origin j to

TABLE 1

Node	Connections to
1	2, 3, 4, 7, 8, 17
2	3, 10, 14, 15, 16
3	8, 12, 16
4	7, 9, 11, 17
5	6, 9, 11, 15, 16, 20
6	7
7	9, 15, 16
8	10, 12, 14, 16, 18
9	12, 20
10	12, 14, 16, 19
11	18, 19, 20
12	13, 15
13	14, 16, 18, 19
14	16, 18, 19
15	16, 17, 19
17	18

TABLE 2

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
−.068	−.089	−.075	−.043	−.055	−.020	−.063	−.097	−.043	−.090	−.040	−.076	−.069	−.098	−.085	−.120	−.042	−.058	−.063	−.023
	1	1					1		1		1	1	1	0	1		1	1	
−.043	.038	.029	−.117	−.108	−.062	−.106	.063	−.118	.076	−.077	.016	.053	.089	−.046	.026	−.054	.021	.027	−.089
1			1	1	1	1		1		1				1		1			1

destination i. The optimal solution of the transportation problem is given in the usual way by integers placed in the lower portion of the squares of the tableau. The partition determned by this solution is $N = \bigcup_{j=1}^{2} N_j$, where

$$N_1 = \{2, 3, 8, 10, 12, 13, 14, 16, 18, 19\},$$
$$N_2 = \{1, 4, 5, 6, 7, 9, 11, 15, 17, 20\}.$$

In this partition nodes 15 and 18 can be interchanged without increasing E_c. When this interchange is made we obtain a partition which agrees with the one found in [11]. It cuts 13 edges.

As a further demonstration of our procedure, we obtain a 7, 7, 6 partition of the graph in Table 1. The transportation tableau for this problem, along with the optimal solution, is given in Table 3. The rows of the tableau are given by $-\sqrt{1/7}u_1$, $-\sqrt{1/7}u_2$, and $-\sqrt{1/6}u_3$, where u_1, u_2, and u_3 are the first three normalized eigenvectors of

TABLE 3

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
−.081	−.106	−.090	−.051	−.066	−.023	−.075	−.116	−.052	−.108	−.048	−.091	−.082	−.117	−.101	−.143	−.050	−.069	−.076	−.027
	1	1					1		1		1		0	1	1				
−.051	.046	.035	−.139	−.129	−.075	−.126	.075	−.141	.091	.093	.020	.063	.106	−.056	.031	−.065	.025	.032	−.106
1			1	1	1	1		1								1			0
.181	.084	.131	.057	−.107	−.009	.084	.054	−.024	−.023	−.148	.010	−.111	−.082	.000	.012	.058	−.088	−.140	−.107
										1		1	1				1	1	1

the adjacency matrix A, oriented so that they are as close to the nonnegative orthant as possible. The partition determined by the solution of this transportation problem is given by $N = \bigcup_{j=1}^{3} N_j$, where

$$N_1 = \{2, 3, 8, 10, 12, 15, 16\},$$
$$N_2 = \{1, 4, 5, 6, 7, 9, 17\}, \tag{5.1}$$
$$N_3 = \{11, 13, 14, 18, 19, 20\}.$$

The partitioned graph is shown in Fig. 1. The partition cuts 23 edges. For this partition, condition (3.1) is satisfied with $j = 1$. We shall therefore investigate the possibility of improving our partition by first obtaining a 13, 7 partition and then a 7, 6 partition as explained in § 3. The transportation tableau corresponding to the 13, 7 partition

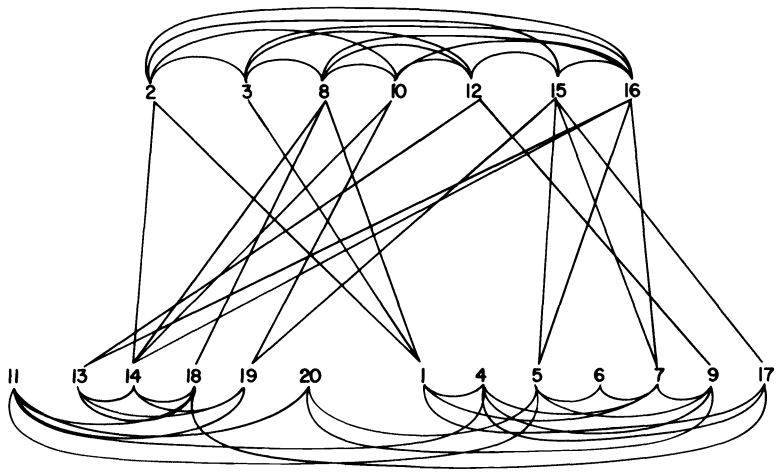


FIG. 1

is shown in Table 4. It gives the partition

$$N_1 = \{1, 2, 3, 8, 10, 12, 13, 14, 15, 16, 17, 18, 19\},$$
$$N_2 = \{4, 5, 6, 7, 9, 11, 20\}$$

which cuts 10 edges. The partitioned graph is shown in Fig. 2.

TABLE 4

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
−.060	−.078	−.066	−.037	−.048	−.017	−.055	−.085	−.038	−.079	−.035	−.067	−.060	−.086	−.074	−.105	−.037	−.050	−.056	−.020
1	1	1					1		1		1	1	1	1	1	1	1	1	
−.051	.046	.035	−.139	−.128	−.075	−.126	.075	−.141	.091	−.093	.020	.063	.106	−.056	.031	−.065	.025	.032	−.106
			1	1	1	1		1		1						0			1

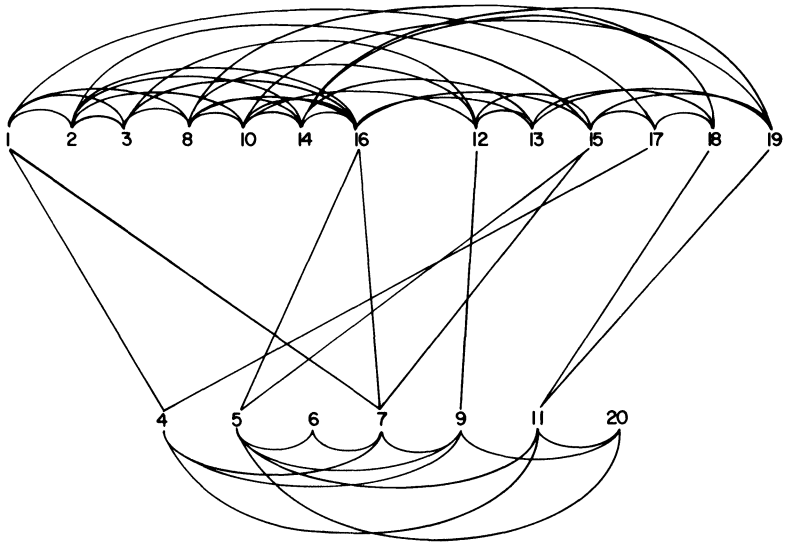


FIG. 2

This partition is probably optimal since its eigenvectors ν_1 and ν_2 are very good approximations of u_1 and u_2 respectively. The 7, 6 partition of N_1 given by our procedure is

$$N_1 = \{1, 2, 3, 8, 10, 14, 16\} \cup \{12, 13, 15, 17, 18, 19\}.$$

It is shown in Fig. 2. This partition cuts 12 edges. Thus

$$N = N_2 \cup \{1, 2, 3, 8, 10, 14, 16\} \cup \{12, 13, 15, 17, 18, 19\}$$

is a 7, 7, 6 partition of our graph which cuts 22 edges. This is an improvement over the partition (5.1) which cuts 23 edges. Both partitions are local minima.

TABLE 5

Node	Connections to
1	7, 12, 13, 14, 15, 16, 17
2	12, 17, 18, 20
3	5, 11, 13, 14, 18, 19, 20
4	6, 9
5	7, 9, 10, 12, 16, 19
6	16, 18, 20
7	8, 9, 11, 16
8	15, 18
9	11, 15, 19
11	14, 17, 18, 20
12	14
13	18, 20
14	16, 18, 20
16	18
17	18
18	20

The second graph in [11] also has 20 nodes. Their connections are given in table 5. We shall partition the nodes into two sets of equal size for comparison with the partition given in [11]. The relevant transportation tableau and the solution of the transportation problem are given in Table 6. The 10, 10 partition obtained from this solution is

$$N = \{2, 3, 6, 11, 13, 14, 15, 16, 17, 18, 20\} \cup \{1, 4, 5, 7, 8, 9, 10, 12, 15, 19\}.$$

It cuts 14 edges. We can interchange nodes 12 and 16 without increasing E_c. Then if we interchange 1 and 6 we arrive at the partition

$$N = \{1, 2, 3, 11, 12, 13, 14, 17, 18, 20\} \cup \{4, 5, 6, 7, 8, 9, 10, 15, 16, 19\}$$

given in [11]. It cuts 13 edges.

TABLE 6

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
−.075	−.055	−.099	−.017	−.065	−.053	−.068	−.036	−.052	−.011	−.100	−.050	−.066	−.104	−.027	−.081	−.059	−.125	−.036	−.100
0	1	1			1					1		1	1		1	1	1		1
−.041	.064	.022	−.032	−.136	.036	−.121	−.031	−.136	−.043	.008	−.025	.054	.035	−.066	.044	.039	.090	−.080	.098
1			1	1		1	1	1	1		1			1				1	

6. Complexity of the algorithm. The algorithm we have described requires three significant calculations for its implementation. First we must determine the k largest eigenvalues and the corresponding eigenvectors of the $n \times n$ adjacency matrix A. We have suggested doing this by the block Lanczos algorithm described in [10]. The principal operation in each step of this algorithm is to form a matrix-vector product of the form Ax. Usually A is very sparse and should be stored in sparse matrix format. Sparse matrix techniques can then be used to form Ax in $O(n)$ operations. We have used the block Lanczos to determine the four largest eigenvalues, and the corresponding eigenvectors, of an adjacency matrix of order $n = 918$. This matrix arises in an actual circuit layout problem. It has an average of 5 nonzeros per row. The block Lanczos algorithm required 73 steps and 5.08 seconds of CPU time on an IBM 370/3033. In general this portion of the algorithm requires $O(n)$ operations.

The next step in our algorithm is to solve a transportation problem involving k origins and n destinations. First we consider the case $k = 2$. This is perhaps the most important case. It is also the easiest to solve. In this case one has a tableau similar to Table 2. m_1 numbers must be selected from the first row and m_2 from the second row, choosing only one number from each column in such a way that the sum of the numbers selected is a minimum. This can be done as follows. Make one pass through the tableau selecting the smaller of the two numbers in each column. Break ties arbitrarily. Suppose this procedure selects m_1' numbers in the first row and m_2' numbers in the second row. If $m_1' = m_1$, then the numbers selected give the solution of the transportation problem. If $m_1' > m_1$, then $m_1' - m_1$ numbers selected in the first row must be dropped in favor of the corresponding numbers in the second row. The numbers to be dropped can be chosen one at a time, always dropping the number that results in the smallest increase to the sum of the numbers selected. If $m_1' < m_1$ a similar adjustment in the numbers selected can be made to arrive at a solution of the transportation problem. Thus when $k = 2$ the transportation problem can be solved by making $O(n)$ comparisons.

Now consider the general case $k > 2$. In the transportation tableau of k rows, duplicate row i m_i times. This gives an $n \times n$ matrix C. Clearly our transportation problem is equivalent to an assignment problem with cost matrix C. Lawler shows in [12, § 4.7] that this problem can be solved in $O(n^3)$ operations. However, this is a worst case analysis, and in practice we have found that the transportation problems are solved very quickly.

The third step of the algorithm involves interchanging pairs of nodes between the various subsets of the partition. In practice we have found this portion of the algorithm to be the most time consuming. Since all pairs of nodes must be considered, this is an $O(n^2)$ operation.

REFERENCES

[1] A. D. FRIEDMAN AND P. R. MENON, *Theory and Design of Switching Circuits*, Bell Telephone Laboratories, and Computer Sciences Press, Rockville, MD, 1975.

[2] R. L. RUSSO, P. H. ODEN AND P. K. WOLFF, SR., *A heuristic procedure for the partitioning and mapping of computer logic blocks to modules*, IEEE Trans. Comput., C-20 (1971), pp. 1455–1462.

[3] H. R. CHARNEY AND D. L. PLATO, *Efficient partitioning of components*, Share/ACM/IEEE Design Automation Workshop, Washington, DC, July 1968.

[4] L. W. COMEAU, *A study of user program optimization in a paging system*, ACM Symposium on Operating System Principles, Gatlinburg, TN, October 1967.

[5] P. J. DENNING, *Virtual memory*, Comput. Surveys, 2 (1970), pp. 153–189.

[6] B. W. KERNIGHAN, *Some Graph Partitioning Problems Related to Program Segmentation*, Ph.D. Thesis, Princeton Univ., Princeton, NJ, 1969.

[7] B. W. KERNIGHAN AND S. LIN, *An efficient heuristic procedure for partitioning graphs*, Bell Systems Tech. J., 49 (1970), pp. 291–307.

[8] A. J. HOFFMAN AND H. W. WIELANDT, *The variation of the spectrum of a normal matrix*, Duke Math. J., 20 (1953), pp. 37–39.

[9] G. HADLEY, *Linear Programming*, Addison-Wesley, Reading, MA, 1962.

[10] J. CULLUM AND W. E. DONATH, *A block Lanczos algorithm for computing the q algebraically largest eigenvalues and a corresponding eigenspace of large, sparse, real symmetric matrices*, in Proc. of the 1974 IEEE Conference on Decision and Control, November 1974, Phoenix, AZ, pp. 505–509.

[11] W. E. DONATH AND A. J. HOFFMAN, *Lower bounds for the partitioning of graphs*, IBM J. Res. Develop., 17 (1973), pp. 420–425.

[12] E. L. LAWLER, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart and Winston, New York, 1976.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0017 $01.00/0

FINITE SOLUTION THEORY FOR COALITIONAL GAMES*

WILLIAM F. LUCAS† AND KAI MICHAELIS†

Abstract. In 1944 von Neumann and Morgenstern introduced a theory of solutions (stable sets) for multiperson cooperative games in characteristic function form. Some special classes of games are known to have solutions which are finite sets. These finite solutions give rise to interesting geometrical structures and basic combinatorial patterns. They have provided new insights into problems in the social sciences and they invite additional interpretations and uses in the mathematical and physical sciences. This paper provides an introduction and survey of finite solution theory.

1. Introduction. In their monumental book *Theory of Games and Economic Behavior* (1944), J. von Neumann and O. Morgenstern [38] introduced the first general model for multiperson coalitional games. Their model consists of four parts: a characteristic function defined on a set of n players, an $(n-1)$-dimensional simplex of imputations, a binary irreflexive relation on this simplex called dominance, and a solution concept which they referred to as solutions. A solution is a subset of the imputations which has certain internal and external stability properties. Such solutions are often referred to currently as stable sets or as von Neumann–Morgenstern solutions, to distinguish them from the many other notions of solution which have since been proposed for such game models.

In general many of the resulting von Neumann–Morgenstern solution sets exhibit some undesirable properties, from both a mathematical and an applied point of view. For many games there are a great number of different solutions and many imputations in each solution. For some games with $n \geqq 10$, solutions may not even exist [18], [19]. There are also games with empty cores (another fundamental and natural solution concept) when $n \geqq 14$ for which solutions do not exist [20], [21]. There are games with $n \geqq 4$ which possess *some* "pathological" sets as solutions [35] and games with $n \geqq 20$ with *only* "pathological" solutions [36]. In general, solutions are difficult to find, and many of them do not seem immediately useful in applications.

On the other hand, there are many large classes of games for which all of (or at least many of) the solutions demonstrate more pleasant theoretical structure and for which new insights of an applied nature have resulted. For example, there is a rich theory of solutions for the "simple" (or voting) games. Symmetric solutions for "symmetric" games exhibit rather intricate and beautiful geometric patterns. Of particular interest are those solution sets which consist of only a finite number of imputations. Although finite solutions do not exist for most games, they are of significant interest when they do appear. They provide a most interesting mathematical structure in their own right as well as new interpretations of an applied nature. Furthermore, finite solutions appear to invite many additional interpretations and uses within mathematics as well as in various physical systems.

The object of this paper is to bring together in one place the bulk of what is known about finite solutions (stable sets), and to present these results in terms of the

* Received by the editors May 17, 1982. This research was supported in part by the National Science Foundation under grants MCS-7728392 and MCS-8102353, and by the Office of Naval Research under contract N00014-75-C-0678, NR 047-094.

† School of Operations Research and Industrial Engineering, College of Engineering, Cornell University, Ithaca, New York, 14853.

modern notation of (0, 1)-normalization. A main purpose is to focus attention on this natural and rather fundamental mathematical structure which arose in applying mathematics to the social sciences, and which now appears ripe for additional interpretations and applications in a variety of other contexts. One new infinite family of finite stable sets is also announced in § 5.3. The original model [38] for characteristic function games is reviewed briefly in § 2, and some special classes of games and necessary notation are introduced in § 3. The known finite stable sets for the n-person games with $n = 3$ and 4 are surveyed in § 4. A few infinite families of finite stable sets for general n are described in § 5. The vast class of games known as extreme games is discussed in § 6. Some suggestions of a rather general nature concerning potential applications appear in § 7.

2. The model. An *n-person game* is given by a set $N = \{1, 2, \cdots, n\}$ of *n players*, and a real valued *characteristic function* $v : 2^N \to \mathbb{R}$ defined on the set 2^N of all subsets (coalitions) of N which has $v(\varnothing) = 0$ for the empty set $\varnothing$. The set of *imputations* (realizable outcomes) for the game (N, v) is defined as the $(n-1)$-dimensional simplex

$$A = \left\{ x \in \mathbb{R}^n : \sum_{i \in N} x_i = v(N) \text{ and } x_i \geqq v(\{i\}) \forall i \in N \right\}$$

where each $x = (x_1, x_2, \cdots, x_n)$ is a feasible distribution of the available wealth $v(N)$ among the individual players. An imputation x *dominates* another one y via the nonempty coalition $S \subset N$ whenever

$$x_i > y_i \quad \forall i \in S$$

and

$$(*) \qquad \sum_{i \in S} x_i \leqq v(S).$$

This is denoted by $x \operatorname{dom}_S y$, or simply as x dom y when some such S exists. For $x \in A$ and $B \subset A$ let

$$\operatorname{Dom} x = \{ y \in A : x \text{ dom } y \}$$

and

$$\operatorname{Dom} B = \bigcup_{x \in B} \operatorname{Dom} x.$$

A *solution* (in the sense of von Neumann–Morgenstern), or *stable set*, for a game is a subset V of A with the properties

$$V \cap \operatorname{Dom} V = \varnothing, \qquad V \cup \operatorname{Dom} V = A$$

which are referred to as *internal* and *external stability*, respectively.

Another solution concept, called the *core* of the game, is given by

$$C = \left\{ x \in A : \sum_{i \in S} x_i \geqq v(S) \forall S \subset N \right\}.$$

It is easy to show for any solution V that $C \subset V$ and $C \cap \operatorname{Dom} V = C \cap \operatorname{Dom} A = \varnothing$.

Many variations and generalizations of the model have been studied. The function v can map coalitions S into sets other than the half spaces given by (*). There is little

change in this theory when the set A is replaced by one of the sets [9]:

$$\bar{A} = \left\{ x \in \mathbb{R}^n : \sum_{i \in N} x_i \leqq v(N) \text{ and } x_i \geqq v(\{i\}) \right\},$$

$$E = \left\{ x \in \mathbb{R}^n : \sum_{i \in N} x_i = v(N) \right\} \quad \text{or} \quad \bar{E} = \left\{ x \in \mathbb{R}^n : \sum_{i \in N} x_i \leqq v(N) \right\}.$$

The set A can also be an arbitrary polyhedron in $\mathbb{R}^n$, or some other type of set. If A is a finite set then one is working in the context of (directed) graph theory. Variations in the definition of dominance can be made, and several other solution concepts have been proposed and analyzed. More generally, an *abstract game* (X, d) consists of an arbitrary set X and an irreflexive binary relation d on X. However, we will not be concerned with these extensions of the classical model in this paper.

The *dominion* $\mathrm{Dom}_S x$ of an imputation x with respect to a given (nonempty) coalition S is either the empty set $\varnothing$ (e.g., when condition (*) fails to hold) or else it is the intersection of A with an open *generalized orthant at x*, i.e., with

$$O_S(x) = \{ y \in \mathbb{R}^n : y_i < x_i \ \forall i \in S \}.$$

A solution V results in a "covering" of $A - V$ by such sets $O_S(x)$ for $x \in V$. The sets $O_S(x)$ may overlap each other as $x \in V$ and $S \subset N$ vary, but they are all disjoint from the closed set V. In the case where a solution V is a finite set, then the resulting sets $O_S(x)$ for $x \in V$ provide a covering of $A - V$ by a finite number of such generalized orthants. If the regions $\mathrm{Dom}_S x$ were viewed as a sort of "directional force field" eminating from x, then a solution V is "held in place" by these resulting domination cones.

3. Some special classes of games. A large number of special classes of games have been defined and studied. The classical model assumed that all games (N, v) were *superadditive*, i.e.,

$$v(S \cup T) \geqq v(S) + v(T) \quad \text{whenever } S \cap T = \varnothing.$$

Dropping this constraint does not alter the theory significantly. However, this condition will be assumed throughout this paper, unless explicitly stated otherwise. Clearly, it is sufficient to consider only *essential* games, ones with

$$v(N) > \sum_{i \in N} v(\{i\}),$$

since A reduces to the empty set $\varnothing$ or else the one point $(v(\{1\}), v(\{2\}), \cdots, v(\{n\}))$ when this latter inequality fails. One can also show that it is sufficient (for von Neumann–Morgenstern solution theory and most of the other known solution concepts) to consider only (0, 1)-*normalized* games, i.e., ones with

$$v(N) = 1 \quad \text{and} \quad v(\{i\}) = 0 \quad \forall i \in N.$$

In this case the set of imputations becomes the unit simplex

$$A = \left\{ x \in \mathbb{R}^n : \sum_{i \in N} x_i = 1, x_i \geqq 0 \ \forall i \in N \right\}.$$

A game is *constant sum* if

$$v(S) + v(N - S) = v(N) \quad \forall S \subset N.$$

This condition reduces the number of independent parameters $v(S)$ by half. A game is *symmetric* whenever

$$|S| = |T| \quad \text{implies} \quad v(S) = v(T)$$

where $s = |S|$ denotes the number of players in the coalition S. A $(0, 1)$-normalized symmetric game is determined by the $n-2$ values $v(2), v(3), \cdots, v(n-1)$ where $v(s) = v(S)$.

Finite solutions arise more frequently in the study of *monotone simple* (or voting) games, i.e., the ones with

$$v(S) = 0 \text{ (losing)} \quad \text{or} \quad v(S) = 1 \text{ (winning)}$$

and

$$S \subset T \quad \text{implies} \quad v(S) \leqq v(T)$$

where S and $T \subset N$. T is called a *minimal winning* coalition if $v(T) = 1$ and $v(S) = 0$ for all $S \subsetneq T$. The lattice $(2^N, \subset)$ is thus "cut" into a set of winning coalitions $\mathcal{W}$ and losing coalitions $\mathcal{L}$ just "below" the set $\mathcal{M}$ of minimal winning coalitions. The *weighted majority* games are the monotone simple games

$$[q; w_1, w_2, \cdots, w_n]$$

where a coalition S is winning if and only if $\sum_{i \in S} w_i \geqq q$. q is called the *quota*. For example, the n-person *direct* (or *simple*) *majority* game is given by

$$[[n/2 + 1]; 1, 1, \cdots, 1]$$

where $[p]$ denotes the largest integer in p.

Many other finite mathematical structures, e.g. projective geometries and block designs, give rise to monotone simple games in a natural way. For example, the seven-point projective plane in which the seven lines correspond to the minimal winning coalitions is a monotone simple game, but is not a weighted majority game [38, 3rd ed., p. 469]. An indication of relations of this type is given in the paper by Bruen [3] and in the references listed there, as well as in the paper by Dubey and Shapley [6].

There are some other major classes of games known as "extreme" games which will be introduced in § 6.

4. Games with small n. The known finite solutions for the games with 3 or 4 players will be discussed in this section. It is assumed that these games are essential, superadditive and $(0, 1)$-normalized. Finite solutions for n-person games with $n \geqq 5$ are presented in the following two sections, and most of these can be viewed as special cases of infinite families of finite solutions.

There is only one 3-person game with a finite solution, and it is $v(\varnothing) = v(1) = v(2) = v(3) = 0$, $v(23) = v(13) = v(12) = v(123) = 1$. (Note that the braces and commas have been deleted from expressions such as $v(\{2, 3\})$.) This particular game is a direct majority game, the only 3-person constant-sum game, and a symmetric game. It has infinitely many solution sets V, but it has only one finite or "symmetric" solution. It consists of the three imputations in the set

$$V^3 = \{(0, \tfrac{1}{2}, \tfrac{1}{2}), (\tfrac{1}{2}, 0, \tfrac{1}{2}), (\tfrac{1}{2}, \tfrac{1}{2}, 0)\}.$$

V^3 is illustrated along with its domination regions in Fig. 1.

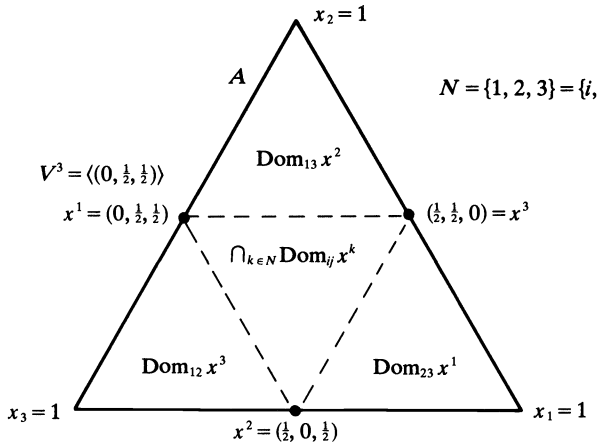


FIG. 1. *The finite symmetric solution* V^3.

Many of the known finite solutions possess total or partial symmetry, and thus it is convenient to introduce the following notation. For any $x \in A$, $B \subset A$ and $S \subset N$ let

$\langle x \rangle = \{y \in A: y \text{ is obtained from } x \text{ by permuting its coordinates}\}$,

$\langle x \rangle_S = \{y \in A: y \text{ is obtained from } x \text{ by permuting its coordinates } x_i \text{ where } i \in S\}$

and

$$\langle B \rangle = \cup_{x \in B} \langle x \rangle.$$

A subset B of A is *symmetric* if $\langle B \rangle = B$. For example, $V^3 = \langle V^3 \rangle = \langle (0, \frac{1}{2}, \frac{1}{2}) \rangle = \langle (0, \frac{1}{2}, \frac{1}{2}) \rangle_N$.

Several finite solutions have been shown to exist for some of the 4-person, constant-sum games. The characteristic function v for these games has $v(S) = 0$ for $|S| = 0$ or 1, $v(S) = 1$ for $|S| = 3$ or 4, and $0 \leqq v(S) = 1 - v(N - S) \leqq 1$ for $|S| = 2$. That is, each such game corresponds to a point $b = (b_1, b_2, b_3)$ in the unit cube U where

$$b_1 = v(14), \quad b_2 = v(24), \quad b_3 = v(34).$$

Using symmetry, it is sufficient to consider only those games corresponding to the points b in this cube which are in the four-sided polyhedron

$$P = \{b \in U: b_1 \leqq b_2 \leqq b_3 \text{ and } b_2 + b_3 \leqq 1\}.$$

P has vertices $(0, 0, 0)$, $(0, \frac{1}{2}, \frac{1}{2})$, $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ and $(0, 0, 1)$, and is illustrated in Fig. 2. Von Neumann and Morgenstern [38] and Mills [26], [27] have determined finite solutions for the four vertices of P, three of the six edges of P, and in a three-dimensional neighborhood in P near the center $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ of the cube U. Two edges and one face of P are known to possess no finite solution. It is known that every three- and four-person general-sum game does possess a solution [1], [38], but most of these known solutions are not finite.

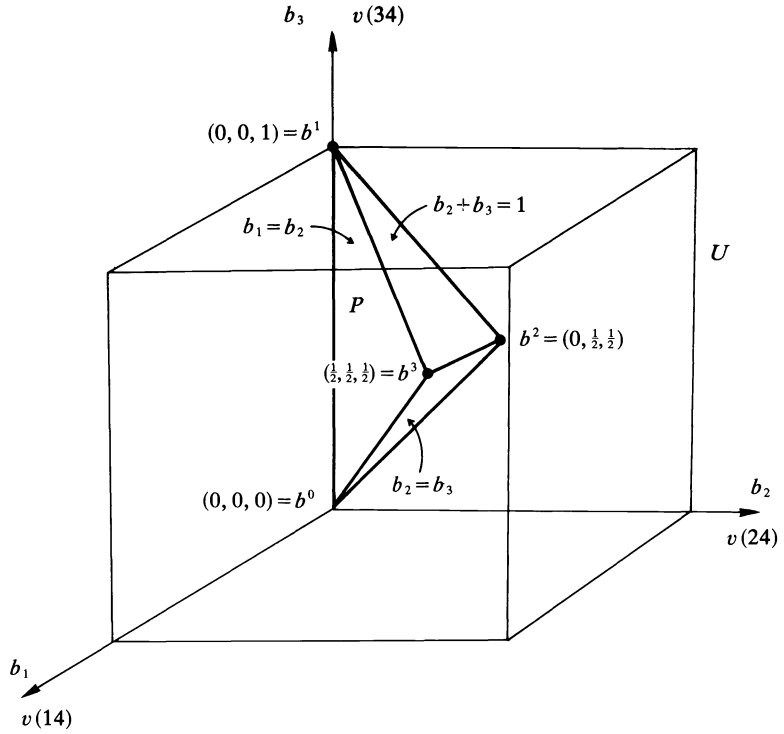


FIG. 2. *The constant-sum four-person games.*

4.1. The vertices of *P*. The three vertices $b^0 = (0, 0, 0)$, $b^1 = (0, 0, 1)$ and $b^2 = (0, \frac{1}{2}, \frac{1}{2})$ of P give rise to the following finite solutions $V^4(b^i)$, $i = 0, 1, 2$, consisting of 3, 4, and 7 imputations, respectively:

$$V^4(b^0) = \langle (0, \tfrac{1}{2}, \tfrac{1}{2}, 0) \rangle_{\{1,2,3\}},$$

$$V^4(b^1) = \{(\tfrac{1}{3}, \tfrac{1}{3}, 0, \tfrac{1}{3})\} \cup \langle (\tfrac{1}{3}, 0, \tfrac{2}{3}, 0) \rangle_{\{1,2,4\}},$$

$$V^4(b^2) = \{(0, \tfrac{1}{2}, \tfrac{1}{2}, 0)\} \cup \langle (0, \tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}) \rangle_{\{1,2,4\}} \cup \langle (0, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}) \rangle_{\{1,3,4\}}.$$

The solution $V^4(b^0)$ is merely the solution V^3 for the three-person game, since player 4 is a "dummy" in the game corresponding to b^0. The vertex $b^3 = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ in P corresponds to the only four-person, constant-sum, symmetric game, and it is known to have at least the following three types of finite solutions of 10, 13, and 13 points, respectively:

$$V_1^4(b^3) = \langle (\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}, 0) \rangle \cup \langle (\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{6}, \tfrac{1}{6}) \rangle,$$

$$V_2^4(b^3(c)) = \{(\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4})\} \cup \langle (\tfrac{3}{8} - c, \tfrac{3}{8} - c, 2c, \tfrac{1}{4}) \rangle \quad \text{where } 0 \leqq c \leqq \tfrac{1}{24},$$

$$V^4(b^3; 1) = \{(\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4})\} \cup \langle (0, \tfrac{1}{4}, \tfrac{3}{8}, \tfrac{3}{8}) \rangle_{\{2,3,4\}}$$
$$\cup \langle (\tfrac{1}{4}, 0, \tfrac{3}{8}, \tfrac{3}{8}) \rangle_{\{2,3,4\}} \cup \langle (\tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{4}, \tfrac{3}{8}) \rangle_{\{2,3,4\}}.$$

The last solution $V_3^4(b^3; 1)$ discriminates against player 1, and is not symmetric. Three similar solutions $V_3^4(b^3; i)$ exist which likewise discriminate against players $i = 2$, 3 and 4, respectively. The solutions $V^4(b^1)$, $V^4(b^2)$, $V_1^4(b^3)$ and $V_3^4(b^3; 1)$ are illustrated in Fig. 3.

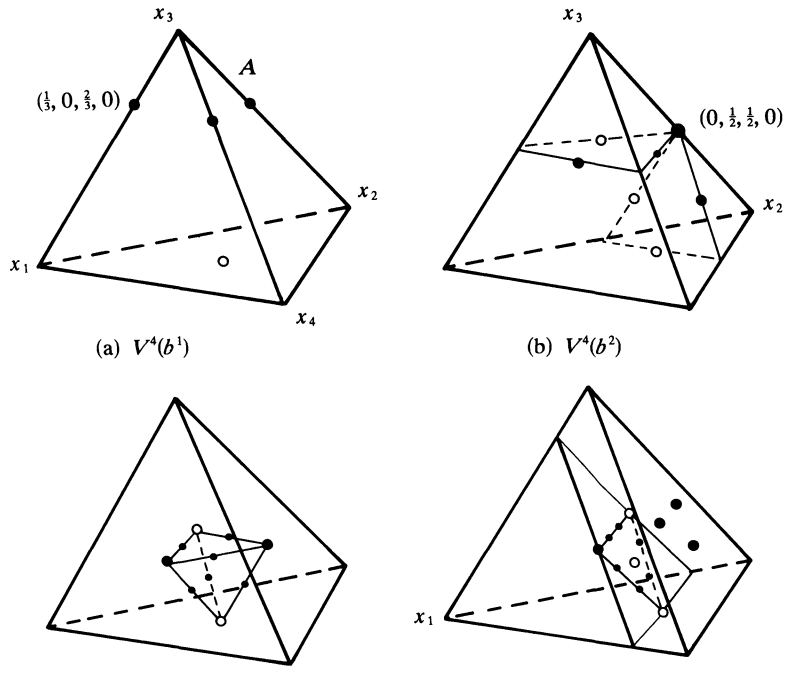


FIG. 3. *Some four-person game solutions.*

4.2. The edges of *P*. Some finite solutions are known for four of the six edges of P.

(i) The "main space" diagonal $b = (z, z, z)$, $0 \leqq z \leqq \frac{1}{2}$, which connects the vertices $b^0 = (0, 0, 0)$ and $b^3 = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ of P has the finite solutions

$$V_z^4\left[b^0, \frac{4}{5}b^3\right] = \left\langle\left(\frac{1}{2}-\frac{z}{4}, \frac{1}{2}-\frac{z}{4}, \frac{z}{2}, 0\right)\right\rangle_{\{1,2,3\}} \cup \left\langle\left(\frac{1}{2}-\frac{z}{4}, \frac{1}{2}-\frac{z}{4}, 0, \frac{z}{2}\right)\right\rangle_{\{1,2,3\}}$$

$$\cup \left\langle\left(\frac{1}{2}-\frac{z}{2}, \frac{1}{2}-\frac{z}{2}, \frac{z}{2}, \frac{z}{2}\right)\right\rangle_{\{1,2,3\}}$$

when $0 \leqq z \leqq \frac{2}{5}$; and

$$V_z^4\left[\frac{4}{5}b^3, b^3\right] = V_z^4\left[b^0, \frac{4}{5}b^3\right] \cup \left\langle\left(\frac{1}{2}-\frac{z}{4}, \frac{1}{2}-\frac{z}{2}, \frac{z}{4}, \frac{z}{2}\right)\right\rangle_{\{1,2,3\}}$$

when $\frac{2}{5} \leqq z \leqq \frac{1}{2}$. These solutions have 9 and 15 imputations, respectively.

(ii) The "main face" diagonal $b = (0, z, z)$, $0 \leqq z \leqq \frac{1}{2}$, which connects the vertices b^0 and $b^2 = (0, \frac{1}{2}, \frac{1}{2})$ of P has the seven-point finite solution

$$V_z^4[b^0, b^2] = \left\{\left(0, \frac{1}{2}, \frac{1}{2}, 0\right)\right\} \cup \left\langle\left(\frac{1}{2}-\frac{z}{2}, \frac{1}{2}, 0, \frac{z}{2}\right)\right\rangle_{\{2,3\}}$$

$$\cup \left\langle\left(\frac{1}{2}-\frac{z}{2}, \frac{1}{2}, \frac{z}{2}, 0\right)\right\rangle_{\{2,3\}} \cup \left\langle\left(\frac{1}{2}-z, \frac{1}{2}, \frac{z}{2}, \frac{z}{2}\right)\right\rangle_{\{2,3\}}$$

which converges to $V^4(b^0)$ and $V^4(b^2)$ as z approaches 0 and $\frac{1}{2}$, respectively. Mills [27] showed that this is the unique finite solution for the *interior* of this edge.

(iii) Consider the other "space" diagonal $(\frac{1}{2}-z, \frac{1}{2}-z, \frac{1}{2}+z)$, $0 \leqq z \leqq \frac{1}{2}$, which joins the vertices b^3 and $b^1 = (0, 0, 1)$ of P. For $0 \leqq z < \frac{1}{18}$, there is the solution $V^4(R(e))$ given below. Von Neumann and Morgenstern [38] stated that finite solutions exist for the intervals $\frac{1}{18} < z \leqq \frac{1}{6}$ and $\frac{1}{6} \leqq z \leqq \frac{1}{2}$, but they did not explicitly display them. It is not known whether finite solutions exist or not when $z = \frac{1}{18}$.

(iv) The "interior" edge $b = (z, \frac{1}{2}, \frac{1}{2})$, $0 \leqq z \leqq \frac{1}{2}$, which joins the vertices b^2 and b^3 of P has the finite solution $V^4(R(e))$ given below when $\frac{2}{5} < z \leqq \frac{1}{2}$. No finite solution is known for $0 < z < \frac{2}{5}$.

(v) and (vi) Mills [26], [27] proved that no finite solution can exist for games corresponding to the interiors of the other two edges of P, i.e., $b = (0, 0, 2z)$ or $(0, \frac{1}{2}-z, \frac{1}{2}+z)$ for $0 < z \leqq \frac{1}{2}$.

4.3. The faces of *P*. Mills [27] proved that there exist no finite solutions on the face of P with $b_1 = 0$, except for the edge $b = (0, z, z)$ where $0 \leqq z \leqq \frac{1}{2}$ and the vertex $b^1 = (0, 0, 1)$, which were covered above. There are no published results about finite solutions for the interiors of the other three faces of P, except where these faces meet the solid region discussed in the next section.

4.4. A neighborhood of b^3. Von Neumann and Morgenstern [38, 3rd ed., pp. 321–329] showed the existence of finite stable sets in a three-dimensional region R in P located near the center point $b^3 = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ of the unit cube U. For any $b \in P$ let

$$u_1(b) = (-1 + b_1 + b_2 + b_3)/2,$$
$$u_2(b) = (1 - b_1 + b_2 - b_3)/2,$$
$$u_3(b) = (1 + b_1 - b_2 - b_3)/2,$$
$$u_4(b) = (1 - b_1 - b_2 + b_3)/2,$$

and then for $i \in N = \{1, 2, 3, 4\}$ define

$$\underline{u}(b) = \min_i u_i(b) \quad \text{and} \quad \bar{u}(b) = \max_i u_i(b).$$

Whenever $b \in P$ and $\frac{2}{3}\bar{u} < 2e \leqq \underline{u}$ then there is the finite solution

$$V^4(R(e)) = \left\{ x \in A : x_i = \begin{cases} u_i + e & \text{if } y_i = \frac{3}{8} \\ u_i & \text{if } y_i = \frac{1}{4} \\ u_i - 2e & \text{if } y_i = 0 \end{cases} \text{ where } y \in V_2^4(b^3(0)) \right\}$$

which has 13 points and is somewhat similar to $V_2^4(b^3(c))$ presented above when $c = 0$. One can prove that the required bounds on the parameter e are satisfied when b is near b^3, e.g., one can pick the region to be

$$R = \{b \in P : 5b_1 + 5b_2 + b_3 > 5\}.$$

5. Infinite families of solutions. For n-person games with $n > 4$ there is relatively sparse knowledge about the existence of finite stable sets. However, there are four types of families of finite solutions for infinitely many values of n which have been discovered. The fourth family, the extreme games of § 6, actually includes the other three families. The three special cases are nevertheless of significant interest in their own right and are discussed briefly in this section.

5.1. The main simple solution. Von Neumann and Morgenstern [38, pp. 431–445] found a finite solution, for each "homogeneous" weighted majority game. A

weighted majority game $[q; w_1, w_2, \cdots, w_n]$ may have infinitely many sets of (nonintegers) weights w_i (even when normalized, e.g. $\sum_{i\in N} w_i = 1$) which give rise to the same simple game, i.e., the same set $\mathcal{M}$ of minimal winning coalitions. The game $[q; w_1, w_2, \cdots, w_n]$ is called a *homogeneous* weighted majority game whenever there exist weights w_i such that the sums $\sum_{i\in S} w_i$ are all equal if and only if $S \in \mathcal{M}$. This occurs when the "overdetermined" system of linear equations $\sum_{i\in S} x_i = 1$ for all $S \in \mathcal{M}$ can be solved. The resulting finite solution relates to the weights in a natural way.

There are four constant-sum, weighted majority games with five persons, and these are all homogeneous. Their main simple solutions are as follows.

(i) $[3; 1, 1, 1, 1, 1]$. This 5-person direct majority game has the *unique symmetric* finite solution

$$V_D^5 = \langle(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}, 0, 0)\rangle$$

consisting of 10 imputations.

(ii) $[4; 1, 1, 1, 2, 2]$ has the seven-point solution

$$V_{h,1}^5 = \{(0, 0, 0, \tfrac{1}{2}, \tfrac{1}{2})\} \cup \langle(\tfrac{1}{4}, \tfrac{1}{4}, 0, \tfrac{1}{2}, 0)\rangle_{\{1,2,3\}} \cup \langle(\tfrac{1}{4}, \tfrac{1}{4}, 0, 0, \tfrac{1}{2})\rangle_{\{1,2,3\}}.$$

(iii) $[5; 1, 1, 2, 2, 3]$ has the five-point solution

$$V_{h,2}^5 = \{(\tfrac{1}{5}, \tfrac{1}{5}, 0, 0, \tfrac{3}{5})\} \cup \langle(0, 0, \tfrac{2}{5}, 0, \tfrac{3}{5})\rangle_{\{3,4\}} \cup \langle(\tfrac{1}{5}, 0, \tfrac{2}{5}, \tfrac{2}{5}, 0)\rangle_{\{1,2\}}.$$

(iv) $[4; 1, 1, 1, 1, 3]$ has the seven-point solution

$$V_{h,3}^5 = \{(\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, 0)\} \cup \langle(\tfrac{1}{4}, 0, 0, 0, \tfrac{3}{4})\rangle_{\{1,2,3,4\}}.$$

5.2. Some symmetric solutions. Bott [2] described a unique symmetric solution $V_B^n(q)$ for each (simple, symmetric, homogeneous) weighted majority game $[q; 1, 1, \cdots, 1]$ for $[(n/2)+1] \leqq q \leqq n$. These solutions give rise to quite interesting geometric structures. However, these are finite solutions only in the case when n is odd and $q = (n+1)/2$, e.g., $V_B^5(3) = V_D^5$ (given above) when $n = 5$. A "symmetric type" solution which is a generalization of the Bott solution $V_B^n(q)$ for quota $q = n-1$ is given in Lucas [17] for games which need not be simple or symmetric.

Muto [29] described a certain class of unique symmetric solutions which are finite. However, the games in this class are *not* in general superadditive. Consider the symmetric n-person games with

$$\begin{aligned} v(s) &= 0 && \text{if } s < k, \\ v(s) &\geqq k/(n-k+1) && \text{if } s = k, \\ v(s) &\leqq sv(k)/k && \text{if } k < s < n \end{aligned}$$

for some k with $2 \leqq k \leqq (n+1)/2$ and where $s = 0, 1, 2, \cdots, n$. The unique symmetric solution is

$$V_M^n(k) = \langle(1/(n-k+1), \cdots, 1/(n-k+1), 0, \cdots, 0)\rangle$$

where each imputation has $n-k+1$ nonzero coordinates. For n odd and $k = (n+1)/2$ this is the Bott solution $V_B^n((n+1)/2)$. In the case where $n = mk-1$ for an integer m and

$$v(s) = \begin{cases} 0 & \text{if } s < k, \\ \dfrac{j}{(m-1)} & \text{if } jk \leqq s < (j+1)k \quad \text{and} \quad 1 \leqq j \leqq m-1, \end{cases}$$

this solution reduces to one given by Gurk and Isbell [13]. This solution has $\binom{n}{n-k+1}$ imputations. For example, when $n=4$, $V_M^4(2)=\langle(\frac{1}{3},\frac{1}{3},\frac{1}{3},0)\rangle$, and if $n=6$ and v satisfies $v(2)\geqq\frac{2}{5}$, $v(3)\leqq\frac{3}{5}$, $v(4)\leqq\frac{4}{5}$ and $v(5)\leqq 1$, then $V_M^6(2)=\langle(\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{5},0)\rangle$.

It should be noted for the games in this section that many of the values $v(S)$ for the "nonvital" coalitions S (i.e., those S not needed in any domination that exists or in the definition of the imputation set A) can often vary over a range of values without altering the resulting solution sets. This is also true for many such nonvital coalitions throughout this paper. It should also be observed that any solution for an n-person game will, by adding 0 coordinates, be a solution of an n'-person game with $n'>n$ which has $n'-n$ "dummy" players, i.e., players who never add any new value by joining a coalition.

5.3. A new family. In 1977, McKelvey and Ordeshook [24] described nonsymmetric solutions $V_0^5(a,\gamma)$ for the direct majority game $[3;1,1,1,1,1]$ which consist of ten imputations of the form

$$(a,a,b,0,0),\quad (0,a,0,b,a),\quad (a,0,0,a,b),\quad (b,0,a,0,a),\quad (a,b,0,0,a),$$

$$(0,0,b,a,a),\quad (a,0,a,b,0),\quad (0,a,a,0,b),\quad (b,a,0,a,0),\quad (0,b,a,a,0)$$

where $2a+b=1$ and $\frac{1}{4}<b<\frac{1}{2}$. (When $a=\frac{1}{4}$ (or $b=\frac{1}{2}$), then the set $V_0^5(\frac{1}{4},\gamma)\cup\langle(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4},0)\rangle$ is a solution.) Each set $V_0^5(a,\gamma)$ is a proper subset of the symmetric set $\langle(a,a,b,0,0)\rangle=W^5$ of 30 points. There are several such solutions $V_0^5(a,\gamma)$ for each value a depending upon the particular selection γ of ten such imputations from the set W^5.

For the 9-person game $[5;1,1,1,1,1,1,1,1,1]$ Michaelis [28] found analogous types of solutions $V_0^9(a,\gamma)$ of 126 imputations each which are subsets of the symmetric set $\langle(a,a,a,a,b,0,0,0,0)\rangle=W^9$ of 630 points and where $4a+b=1$ and $\frac{1}{6}<b<\frac{1}{3}$. (When $a=\frac{1}{6}$ (or $b=\frac{1}{3}$), then $V_0^9(\frac{1}{6},\gamma)\cup\langle(\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},0,0,0)\rangle$ is a solution.) He also proved that there are no such solutions contained in $W^7=\langle(a,a,a,b,0,0,0)\rangle$ for the 7-person direct majority game.

Recently, it has been shown that when n is odd and not of the form 2^p-1, then the game $[(n+1)/2;1,1,\cdots,1]$ has various solutions $V_0^n(a,\gamma)$ of $\binom{n}{(n+1)/2}$ imputations each. These are proper subsets of the symmetric set $\langle(a,\cdots,a,b,0,\cdots,0)\rangle=W^n$ of $(n+1)/2(\binom{n}{(n+1)/2})$ points and have $(n-1)a/2+b=1$ and $2/(n+3)<b<4/(n+3)$ (or $2/(n+3)<a<2(n+1)/(n+3)(n-1)$). No solution of this form can exist when n is of the form 2^p-1, because $\binom{n}{(n+1)/2}$ is then odd and the following characterization is impossible to achieve.

One can characterize these solutions $V_0^n(a,\gamma)$ as the subsets of $\langle(a,\cdots a,b,0,\cdots,0)\rangle=W^n$ which are *complete* in the sense that for each $S\subset N$ with $|S|=(n-1)/2$ there is a unique imputation x with $x_i=0$ for all $i\in S$, and *complementary* in the sense that if $x\in V_0^n(a,\gamma)$ then x' is also in this solution where $x_i'=a$ when $x_i=0$ and $x_i'=0$ when $x_i=a$. The detailed proof of this characterization and of the existence of such sets appears in Lucas, Michaelis, Muto and Rabie [22], [23].

6. Extreme games. Many of the particular games mentioned so far belong to a special class of games known as "extreme" games. This class of games contains most of the games for which finite solutions have been determined, and it provides a useful scheme for studying games with finite solutions. Only a brief introduction to extreme games is presented here. A more detailed exposition is given in the monograph on this topic by Rosenmüller [33].

We will restrict our considerations to essential, superadditive and constant-sum games in (0, 1)-normalization. Any such n-person game is determined by the $2^{n-1}-n-$

1 parameters $v(S)$, and the set of all such games forms a polyhedral subset P of the unit cube U of $2^{n-1}-n-1$ dimensions. One can define the class of *extreme* games as those games corresponding to the extreme points of P.

One also classifies games (whether extreme or not) according to the number of distinct values the characteristic function can take on. This approach extends the notion of weighted majority games to the case of several quotas q_j and coalitional values $v(S)$ as follows. Consider any n-person game for which the players have *integer weights* $w_1, w_2, \cdots, w_n$; there are E quotas $q_1 \leqq q_2 \leqq \cdots \leqq q_E$, and the characteristic function v is of the form

$$(**) \qquad v(S) = \begin{cases} 0 & \text{if } w(S) < q_1, \\ d_j & \text{if } q_j \leqq w(S) < q_{j+1}, \\ 1 & \text{if } q_E \leqq w(S) \end{cases}$$

where $w(S) = \sum_{i \in S} w_i$. Given the vector $w = (w_1, w_2, \cdots, w_n)$ define its *restriction* w^S to S by

$$w_i^S = \begin{cases} w_i & \text{if } i \in S, \\ 0 & \text{if } i \notin S. \end{cases}$$

Let

$$H(q_j) = \{S \subset N : w(S) = q_j\}.$$

A weight vector w is called *homogeneous* for the quota vector $q = (q_1, q_2, \cdots, q_E)$, and denoted by w hom q, whenever $w(S) \geqq q_j$ implies that there is a $T \subset S$ such that $w(T) = q_j$. Note that a game of the form (**) need not in general be superadditive unless additional constraints are placed on the w_i, d_j and q_j.

A game of form (**) is a (simple) weighted majority game if $E = 1$. (**) describes a symmetric game if each $w_i = 1$. It is convenient to classify *extreme* games of type (**) according to the numbers $E+1$ of values taken on by v. Finite solutions are known to exist for some extreme games of form (**) for various values of $E+1$ from 2 through 4. A brief indication of some of these results follows. This category includes most of the games described in §§ 4 and 5. See Rosenmüller [33] for a more extensive treatment of extreme games.

6.1. Extreme games with two values. If a constant-sum n-person game of form (**) with $E+1=2$ is homogeneous, then it has the *main simple solution*

$$V_h^n(q) = \{w^S/q : S \in H(q)\}$$

of von Neumann and Morgenstern [38] where $q = q_1$ is the quota. The four constant-sum, simple, 5-person games are homogeneous weighted majority games, and their main simple solutions were given in § 5.1. There are 23 simple, constant-sum, 6-person games. Eight of these are homogeneous weighted majority games. For seven other games of this type Gurk and Isbell [13] determined finite solutions by generalizing the main simple solution to consider more general sets H' of coalitions instead of $H(q)$. For the eight remaining games in this class no finite solutions have been determined.

6.2. Extreme games with three values. When the characteristic function in (**) takes on three values (i.e., when $E=2$), then the conditions for extremality and the determination of finite solutions become more involved. The constant-sum

requirement determines $d_1 = \frac{1}{2}$ and for the integers w_i that $q_1 + q_2 = w(N) + 1$. The resulting game will be extreme if in addition

$$[(w(N)+2)/3] \leqq q_1 \leqq [(w(N)-1)/2],$$

$w_i < q_2 - q_1$, $i = 1, 2, \cdots, n$, and w hom q_1, w hom q_2, and w hom $(q_2 - q_1)$ are satisfied. This is not a complete characterization of extreme games (with three values), as there are extreme games which cannot be described by weights and quotas as in (**). Let $\alpha, \beta \in \mathbb{R}$ and let

$$V_E^n(q_j; \alpha, \beta) = \{(w^R + \alpha w^S)/q_E : R \in H(q_j), S \in H(\beta), R \cap S = \varnothing\}.$$

For two smaller intervals of q_1 and additional homogeneity requirements on w, Rosenmüller [33] extended the main simple solution. He showed that if $[(w(N)+4)/3] \leqq q_1 \leqq [3(w(N)+1)/7]$ and also if w hom $(2q_1 - q_2)$ and w hom $(3q_2 - 4q_1)$, then $V_{R,1}^n = V_h^n(q_2) \cup V_2^n(q_1; \frac{1}{2}, 2(q_2 - q_1))$ is a solution. Also if $[(w+4)/3] \leqq q_1 \leqq [(w-1)/2]$ and w hom $(q_2/2)$ and w hom $(q_1 - q_2/2)$,

$$V_{R,2}^n = V_h^n(q_2) \cup V_2^n(q_1; 2(q_2 - q_1)/q_2, q_2/2)$$

is a finite solution. For the special case of symmetric games, i.e. games with $w_1 = w_2 = \cdots = w_n = 1$, Griesmer [11] showed that if $q_1 = (N+1)/3$ is an integer and $d_1 = \frac{1}{2}$, then $q_2 = 2q_1$ and the game given by (**) is extreme and has $V_h^n(q_2)$ as a finite solution.

There are no 4-person extreme games with three values. Gurk [12] determined all eight extreme 5-person games with three values and stated that he found finite solutions for all but two of these games. One of these eight extreme games is also a symmetric game.

6.3. Extreme games with four values. The conditions on the characteristic function (**) become more complex when it takes on four values (i.e., $E = 3$). To be constant-sum $d_1 = \frac{1}{3}$, $d_2 = \frac{2}{3}$ and $q_1 + q_3 = w(N) + 1$, $q_2 = (w(N)+1)/2$. The resulting game will be extreme if the following conditions hold:

$$\begin{aligned}
&\tfrac{1}{2}q_2 \leqq q_1 \leqq \tfrac{2}{3}q_2, \\
&w_i < q_3 - q_2, \qquad w_i < q_2 - q_1 \quad \text{for } i = 1, 2, \cdots, n, \\
&w \text{ hom } q_j \quad \text{for } j = 1, 2, 3, \\
&w \text{ hom } (q_3 - q_2), \qquad w \text{ hom } (q_2 - q_1).
\end{aligned}$$

Again, this is not a complete characterization of extreme games with four values. For symmetric games (all $w_i = 1$) with an odd number of 7 or more players which satisfy the above conditions, Muto [30] obtained finite solutions. If $3q_2/5 < q_1 \leqq 5q_2/8$, then

$$V_M^n = V_h^n(q_3) \cup \left\{ w^R/q_3 + 2w^S/3q_3 + \frac{(10q_1 - 3w(N))w^T}{(18q_1 - 6q_3)q_3} : \right.$$

$$\left. R \in H(q_2), S \in H(3q_2 - 4q_1), S \cap R = \varnothing, T = N - (S \cup R) \right\}$$

is a finite solution. He also obtained a finite solution with a similar structure for $q_2/2 \leqq q_1 \leqq 2q_2/5$.

Rosenmüller [33] showed that when $q_1 = 3q_2/5$ and $w(N)+1$ is a multiple of 10, then

$$V_R^n = V_h^n(q_3) \cup V_3^n(q_2; 2/3, 3(q_3 - q_2)/2 \cup V_3^n(q_1; 2/3, 3(q_3 - q_1)/2)$$

is a finite solution for a game satisfying the conditions above. Actually, his theorem is more general in that it allows games given by (**) to also have more than four values, and as long as $q_j = (j(E-1)+1)(w(N)+1)/(E^2+1)$, w hom q_1 and $w(N)+1$ is a multiple of E^2+1,

$$V_R^n = V_h^n(q_E) \cup \left[\bigcup_{j=1}^{E-1} V_E^n(q_j; (E-1)/E, E(q_E - q_j)/(E-1)) \right]$$

will be a finite solution and the resulting game will be an extreme game with $E+1$ values.

7. Potential applications. When finite solutions do exist they provide a very interesting mathematical structure. They usually exhibit some pleasant symmetries as well as some intricate combinatorial properties. Recall that a finite solution V gives a covering of the imputation simplex A by means of a finite number of overlapping, open generalized orthants $O_S(x)$ whose complement in A is just the set V. (A can also be replaced in this theory by the n-dimensional simplex $\bar{A}$, the "subspace" E, or the half space $\bar{E}$.) It would appear as though such finite solutions give rise to quite natural and perhaps rather basic structures and interrelationships which could prove of significant mathematical interest from a purely theoretical point of view. They form a certain type of geometry of points and "space" filling cones emanating from these points. In any case, such solutions have already proved to be of serious interest to applications arising in the social sciences. In addition, finite solutions seem to suggest various additional interpretations within applied mathematics as well as in the physical sciences.

Much of classical geometry deals with points, lines and subspaces, and relations between these. These concepts are highly linear ones and "lower-dimensional" in nature, whereas many areas in contemporary mathematics such as discrete optimization make use of more "directional" or "angular" types of notions such as cones. These may be full-dimensional regions, and may display their own geometrical and combinatorial relationships. Many new mathematical concepts, subject areas, abstractions and syntheses are resulting from recent developments in fields such as optimization theory and combinatorics. Ideas such as the blocking polyhedra of Fulkerson [8] or the corner polyhedra of Gomory and Johnson [10] are examples. There are recent results on the rigidity of polyhedra and bar systems such as flexible (nonconvex) polyhedra by Connelly [5] and related work by Whitley [39] and B. Roth whose rigidity, or lack thereof, seems to invite additional interpretations in terms of angular notions; this may be particularly true for the analogous problems in higher dimensions. Attempts to give geometrical descriptions to mathematical entities such as spinors [4], [32] or twistors [7], [14] usually rely heavily upon directions and angles, at least in lower dimensional cases. Finite solutions may be one of several possible ways to construct useful geometric systems based more on angular notions than linear subspaces.

The combinatorial structure of some finite solutions often leads to immediate applications beyond the original game theoretical context. The recently discovered family of solutions described in § 5.3 can be used in statistical designs, for more efficient storage of computer data or in scheduling workers or athletes (e.g., baseball pitchers). This is discussed in another paper by Lucas, Michaelis, Muto and Rabie [22], [23].

There are many physical systems in which small particles are "held in position" in space. Crystals and molecules are examples. Even superfluids such as normal helium

and helium three, He^3, at very low temperatures behave somewhat like a giant molecule or crystal. For example, they flow or rotate with little friction and are slightly magnetic. The forces arising from each "particle" in such structures would appear to be less than uniform in all directions. On the other hand, there is, in many parts of the physical sciences, a strong habit of thinking in terms of central force fields, whereas interactions of elementary particles and the structure of the atom or the nucleus appear to be governed by a variety of discrete variables or rules which appear somewhat combinatorial in nature. One can easily conceive of models where particles with angular force fields (perhaps pulsating as well) bind together in given integral numbers when at close range, and yet repulse each other at larger distances. Such (nonlinear) fields may be angular at short distances and appear as central force fields or take on a probabilistic interpretation at larger distances. Such models seem more natural or straightforward than explanations in terms of "spring-like" forces or being enclosed in a "tough skin". There exist models which view current *elementary* particles as combinations of other objects, as purely combinatorial-type structures in twistor space [7], [14], or as related to the notion of monopoles [15]. Evidence on the existence of fractional electrical charge [16] would appear to support these former views. In light of this situation, it does not seem too unreasonable to suggest that the employing of additional combinatorial ideas or the more explicit shift in thinking to more angular notions in the study of some physical systems may be worthwhile.

Speculations about the value of a mathematical theory "in search of an application" should be taken with due caution. Plato and Kepler were badly misled by trying to fit the real world to the beautiful result on the existence of precisely five regular (Platonic) polyhedra [34]. On the other hand, this classical result has provided a guide to some useful discoveries such as the construction of the three synthetic hydrocarbons called the cubane, the tetrahedrane, and recently the dodecahedrane [31]. The theory of solutions (whether finite or not) has already made a contribution to mathematical modeling in the direction of the social sciences, and the dominance relation used in this theory will clearly be a very essential ingredient in many more models concerning multiperson interactions and coalitional behavior. The theory of finite stable sets which grew out of such an attempt to model social situations does provide an interesting mathematical system which also seems to invite additional applications in the direction of the physical sciences.

REFERENCES

[1] O. N. BONDAREVA, *The solution of classical cooperative four-person games with nonempty core (the general case)*, Vestnik Leningrad. Univ. Mat. Mekh. Astronom., 4 (1979), pp. 14–19 and 121 (see also pp. 42–47 and 52–60). (In Russian.)

[2] R. BOTT, *Symmetric solutions to majority games*, in Contributions to the Theory of Games, Vol. II, H. W. Kuhn and A. W. Tucker, eds., Annals of Math. Studies, 28, Princeton Univ. Press, Princeton, NJ, 1953, pp. 319–323.

[3] A. BRUEN, *Blocking sets in finite projective planes*, SIAM J. Appl. Math., 21 (1971), pp. 380–392.

[4] ELIE CARTAN, *The Theory of Spinors*, Dover, New York, 1981. (Original notes in French, 1937.)

[5] ROBERT CONNELLY, *A counterexample to the rigidity conjecture for polyhedra*, Publ. Math., 47, Institut des Hautes Etudes Scientifiques, Paris, 1978.

[6] P. DUBEY AND L. S. SHAPLEY, *Mathematical properties of the Banzhaf power index*, Math. Oper. Res., 4 (1979), pp. 99–131.

[7] R. L. FORWARD, *Spinning New Realities*, Science 80 (Dec. 1980), 40–49.

[8] D. RAY FULKERSON, *Blocking and anti-blocking pairs of polyhedra*, Math. Programming, 1 (1971), pp. 168–194.

[9] D. B. GILLIES, *Solutions to general non-zero-sum games*, in [37], pp. 47–85.

[10] R. E. GOMORY AND E. L. JOHNSON, *Some continuous functions related to corner polyhedra, I and II*, Math. Programming, 3 (1972), pp. 23–85 and 359–389.
[11] J. H. GRIESMER, *Extreme games with three values*, in [37], pp. 189–212.
[12] H. M. GURK, *Five-person, constant-sum, extreme games*, in [37], pp. 179–188.
[13] H. M. GURK AND J. R. ISBELL, *Simple solutions*, in [37], pp. 247–265.
[14] L. P. HUGHSTON AND R. S. WARD, eds., *Advances in Twistor Theory*, Pitman, London, 1979.
[15] ARTHUR JAFFE AND CLIFFORD TAUBES, *Vertices and Monopoles: Structure of Static Gauge Theories*, Birkhauser, Boston, 1980.
[16] G. S. LARUE, J. D. PHILLIPS AND W. M. FAIRBANKS, *Observation of fractional charge of* (1/3)e *on matter*, Phys. Rev. Lett., 46 (1981), pp. 967–970.
[17] W. F. LUCAS, *n-person games with only* 1, $n-1$, *and n-person coalitions*, Z. Wahrsch. Verw. Geb., 6 (1966), pp. 287–292.
[18] ———, *A game with no solution*, Bull. Amer. Math. Soc., 74 (1968), pp. 237–239.
[19] ———, *The proof that a game may have no solution*, Trans. Amer. Math. Soc., 137 (1969), pp. 219–229.
[20] W. F. LUCAS AND M. RABIE, *Existence theorems in game theory*, School of Operations Research and Industrial Engineering, Tech. Report 473, Cornell Univ., Ithaca, NY, July 1980.
[21] ———, *Games with no solutions and empty cores*, Math. Oper. Res., 7 (1982) to appear.
[22] W. F. LUCAS, K. MICHAELIS, S. MUTO AND M. RABIE, *A new family of finite solutions*, Intern. J. Game Theory, 11 (1982), to appear.
[23] ———, *Detailed proofs for a family of finite solutions*, School of Operations Research and Industrial Engineering, Tech. Report 523, Cornell Univ., Ithaca, NY, 1981.
[24] R. D. MCKELVEY AND P. C. ORDESHOOK, *An undiscovered von-Neumann–Morgenstern solution for the* (5, 3) *majority rule game*, Internat. J. Game Theory, 6 (1977), pp. 33–34.
[25] N. D. MERMIN AND D. M. LEE, *Superfluid helium* 3, Scientific American, 235 (Dec., 1976), pp. 56–71.
[26] W. H. MILLS, *The four person game-edge of the cube*, Ann. of Math., 59 (1954), pp. 367–378.
[27] ———, *The four person game—finite solutions on the face of the cube*, in [37], pp. 125–143.
[28] K. MICHAELIS, *A survey of finite stable sets for cooperative games*, M.S. Thesis in Operations Research, Cornell Univ., Ithaca, NY, 1981.
[29] S. MUTO, *Stable sets for symmetric, n-person, cooperative games*, Tech. Report 387, School of Operations Research and Industrial Engineering, Cornell Univ., Ithaca, NY, 1978.
[30] ———, *Symmetric solutions for symmetric, constant-sum, extreme games with four values*, Internat. J. Game Theory, 8 (1979), pp. 115–123.
[31] L. A. PAQUETTE, D. W. BALOGH, R. USHA, D. KOUNTZ AND G. G. CHRISTOPH, *Crystal and molecular structure of a pentagonal dodecahedron*, Science, 211 (1981), pp. 575–576.
[32] W. T. PAYNE, *Elementary spinor theory* and *A geometric approach to nonrelativistic spin theory*, Amer. J. Phys., 20 (1952), pp. 253–262; and 21 (1953), pp. 621–628.
[33] J. ROSENMÜLLER, *Extreme Games and Their Solutions*, Lecture Notes in Economics and Mathematical Systems, 145, Springer, New York, 1977.
[34] CARL SAGAN, *Cosmos*, Random House, New York, 1980.
[35] L. S. SHAPLEY, *A solution with an arbitrary closed component*, in [37], pp. 87–93.
[36] ———, *Notes on n-person games* VIII: *A game with infinitely "flaky" solutions*, unpublished manuscript, 1968.
[37] A. W. TUCKER AND R. D. LUCE, eds., *Contribution to the Theory of Games*, Vol. IV, Annals of Math. Studies, 40, Princeton Univ. Pres , Princeton, NJ, 1959.
[38] J. VON NEUMANN AND O. MORGENSTERN, *Theory of Games and Economic Behavior*, Princeton Univ. Press, Princeton, NJ, 1944; 3rd ed., 1953.
[39] WALTER WHITELEY, *Motions, stresses and projected polyhedra*, draft paper, March 1981.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0018 $01.00/0

DOUBLE SEMIORDERS AND DOUBLE INDIFFERENCE GRAPHS*

MARGARET B. COZZENS† AND FRED. S. ROBERTS‡

Abstract. The notion of semiorder was introduced by Luce in 1956 as a model for preference in the situation where indifference judgments are nontransitive. The notion of indifference graph was introduced by Roberts in 1968 as a model for nontransitive indifference. Motivated by problems of measurement and seriation in the social sciences and by frequency assignment problems in communications, we discuss generalizations called double semiorders and double indifference graphs. Semiorders are exactly the binary relations (A, P) such that there is a real-valued function f on A satisfying xPy iff $f(x)>f(y)+\delta$, where δ is a fixed positive number. Indifference graphs are exactly the graphs (V, E) such that there is a real-valued function f on V satisfying $\{x, y\}\in E$ iff $|f(x)-f(y)|\leqq\delta$. Suppose $\delta_1>\delta_2>0$. We present conditions on a pair of binary relations (A, P_1) and (A, P_2) necessary and sufficient for the existence of a real-valued function f on A satisfying xP_iy iff $f(x)>f(y)+\delta_i$, $i=1, 2$. These lead to conditions on (V, E_1) and (V, E_2) necessary and sufficient for the existence of a real-valued function f on V satisfying $\{x, y\}\in E_i$ iff $|f(x)-f(y)|\leqq\delta_i$, $i=1, 2$.

1. Introduction.[1] The notion of semiorder was introduced by Luce [1956] as a model for an individual's preferences when, due to the existence of thresholds, indifference is not transitive. The analogous notion for indifference, called an indifference graph, was introduced by Roberts [1968], [1969]. In this paper, we study generalizations of semiorders and indifference graphs which arise from problems in measurement and seriation in the social sciences and from problems of frequency assignment in communications, when more than one threshold exists.

Suppose P is a binary relation of preference on a set A. Luce [1956] asked for conditions on (A, P) necessary and sufficient for the existence of a real-valued function f on A so that for all $x, y\in A$,

$$xPy \Leftrightarrow f(x)>f(y)+\delta, \tag{1}$$

where δ is a fixed positive number. If such a function f exists, x is preferred to y if and only if its measure $f(x)$ is "sufficiently larger" than $f(y)$, where "sufficiently larger" is measured by a threshold δ. In turn, if indifference I is defined on A by $xIy \Leftrightarrow \sim xPy \;\&\; \sim yPx$, Roberts [1968], [1969] asked for conditions on (A, I) necessary and sufficient for the existence of a real-valued function f on A so that for all $x, y\in A$,

$$xIy \Leftrightarrow |f(x)-f(y)|\leqq\delta. \tag{2}$$

Modifying a notion of Luce, Scott and Suppes [1958] defined a *semiorder* (A, P) as a binary relation satisfying the following conditions for all a, b, c, d in A:

S1: $\sim aPa$,
S2: $aPb \wedge cPd \Rightarrow aPd \vee cPb$,
S3: $aPb \wedge bPc \Rightarrow aPd \vee dPc$.

THEOREM 1 (Scott and Suppes [1958]). *Suppose P is a binary relation on a finite set A, and δ is a positive number. Then there is a real-valued function f on A such that* (1) *is satisfied if and only if (A, P) is a semiorder.*

* Received by the editors June 15, 1981. This work was partially supported by the U.S. Air Force Office of Scientific Research under grant AFOSR-80-0196 to Rutgers University.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115.
‡ Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903.

[1] We adopt the relation-theoretic terminology and notation of Roberts [1979b] and the graph-theoretic terminology and notation of Harary [1969] and Roberts [1976], [1978].

COROLLARY 1.1. *If a binary relation P on a finite set A is representable in the form* (1) *for some positive number δ, then it is representable in the form* (1) *for any positive number δ.*

For alternative proofs of the Scott–Suppes theorem, see Scott [1964], Rabinovitch [1978], and Roberts [1979b].

If a binary relation (A, I) is representable in the form (2), then certainly I is reflexive and symmetric. We may define a graph $G=(V, E)$ by taking $V=A$ and $\{x, y\}\in E$ iff xIy. This is an undirected graph with a loop at each vertex. All our graphs will have this property and will also have the property that V is finite. However, we shall always omit the loops in our diagrams. We then ask whether there is a real-valued function f on V so that for all $x, y \in V$,

$$\{x, y\}\in E \Leftrightarrow |f(x)-f(y)|\leqq\delta. \tag{3}$$

We say G is an *indifference graph* if there is a function f satisfying (3).

COROLLARY 1.2 (Roberts [1969]). *A graph G is an indifference graph if and only if the edges of the complementary graph $\tilde{G}$ can be oriented so that the resulting binary relation defines a semiorder.*

Roberts [1969] also presents several other characterizations of indifference graphs, including a forbidden subgraph characterization.

In this paper, we shall consider pairs of preference relations and indifference relations on the same set. We shall seek representing functions f satisfying analogues of (1) and (3) where there is more than one threshold. Specifically, suppose P_1 and P_2 are binary relations on A, (V, E_1) and (V, E_2) are graphs and $\delta_1>\delta_2$ are positive numbers. Then we seek necessary and sufficient conditions for the existence of real-valued functions f on A (respectively, V) so that for all x, y in A (respectively, V),

$$xP_1y \Leftrightarrow f(x)>f(y)+\delta_1 \quad \text{and} \quad xP_2y \Leftrightarrow f(x)>f(y)+\delta_2 \tag{4}$$

and

$$\{x, y\}\in E_1 \Leftrightarrow |f(x)-f(y)|\leqq\delta_1 \quad \text{and} \quad \{x, y\}\in E_2 \Leftrightarrow |f(x)-f(y)|\leqq\delta_2. \tag{5}$$

We shall explain applications of these representations below. However, we can think of the two thresholds as distinguishing two levels of preference or indifference, strong and weak, or two levels of interference, strong and weak, etc.

The paper is organized as follows. Sections 2, 3, 4 and 5 present applications of and motivations for the representations (4) and (5), and § 6 presents a precise formulation of our questions. Section 7 presents a general method for solving such representation problems, §§ 8 and 9 present our specific solutions, and § 10 presents open problems.

2. The channel assignment problem. One of a number of related frequency assignment problems in communications theory is the *channel assignment problem*, which has been studied by Zoellner and Beall [1977] and by Hale [1980]. It can be defined in general terms as follows. Suppose we have a set V of transmitters, each of which is to be assigned a discrete frequency or *channel*, a positive integer, on which to transmit its signal. Suppose there are k different levels of interference between transmitters. A typical criterion is that u and v interfere at level i if and only if $d(u, v)\leqq\delta_i$, where δ_i is a positive real number and $d(u, v)$ is some appropriate measure of the distance between the transmitters u and v. Suppose $T(i)$ is a given set of

disallowed distances between channels, with $\{0\} \subseteq T(1) \subseteq T(2) \subseteq T(3) \subseteq \cdots \subseteq T(k)$. We seek an assignment of a channel $A(v)$ for each v in V so that for all $u, v \in V$,

(6) $\qquad u$ and v interfere at level $i \Rightarrow |A(u) - A(v)| \notin T(i)$.

For UHF television transmission, the distances δ_i and sets $T(i)$ are given as follows: $\delta_1 = 155$, $\delta_2 = 75$, $\delta_3 = 60$, $\delta_4 = 55$, $\delta_5 = 20$, $T(1) = \{0\}$, $T(2) = \{0, 15\}$, $T(3) = \{0, 7, 14, 15\}$, $T(4) = \{0, 1, 7, 14, 15\}$, $T(5) = \{0, 1, 2, 3, 4, 5, 7, 8, 14, 15\}$ (Hale [1980]).

Let G_i be a graph with vertex set V, and an edge between vertices u and v if and only if u and v interfere at level i. If the transmitters are points on a line, and interference is defined by $d(u, v) \leqq \delta_i$, then each G_i is an indifference graph. More generally, if the transmitters are thought of as points in the plane or in 3-space, then the graphs G_i are said to be graphs of (*unit*) *sphericity* at most 2 or at most 3, respectively. See Cozzens [1981] and Havel [1982] for further discussion of this case. The case $k = 2$ gives a pair of graphs G_1 and G_2 so that

(7) $$\{x, y\} \in E(G_1) \Leftrightarrow d(x, y) \leqq \delta_1$$

and

(8) $$\{x, y\} \in E(G_2) \Leftrightarrow d(x, y) \leqq \delta_2.$$

In studying the representation (5), we seek to characterize the pairs of graphs G_1 and G_2 which arise as in (7) and (8), provided distance is between points on a line. The general characterization problem is still open for different metrics d and different values of k.

Hale [1980] points out two important special cases of the general channel assignment problem we have posed. One is where $k = 1$ and $T(1) = \{0\}$. Then a channel assignment satisfying (6) is an ordinary graph coloring, since $|A(u) - A(v)| \neq 0$ is equivalent to $A(u) \neq A(v)$. One often seeks a channel assignment using as small a number of channels as possible, and hence in this special case, a graph coloring using as few colors as possible. Grötschel et al. [1980] have shown that the problem of finding the graph coloring using the fewest colors is solvable by an efficient (polynomial) algorithm for perfect graphs, and hence in particular for indifference graphs. Hsu [1980] also shows this for claw-free perfect graphs, which include indifference graphs. In general, the problem of finding the best coloring is NP-complete. Indeed, James Orlin [personal communication] has shown that it is NP-complete even for graphs of (unit) sphericity at most 2.

A second special case is where $k = 2$ and $T(1) = \{0\}$ and $T(2) = \{0, 1\}$. Then we seek a function $A(v)$ satisfying the following conditions:

(9) $$\{u, v\} \in E(G_1) \Rightarrow A(u) \neq A(v),$$

(10) $$\{u, v\} \in E(G_2) \Rightarrow A(u) \neq A(v) \text{ and } |A(u) - A(v)| \neq 1.$$

i.e., $A(u)$ and $A(v)$ are not adjacent channels. One would like to find channel assignments A satisfying (9) and (10) and using as small a number of channels as possible. A first step in developing procedures for solving this problem is to characterize pairs of graphs arising as in (7) and (8), in particular to characterize pairs of graphs representable in the form (5). This problem we shall solve below.

For recent results on the channel assignment problem, see Cozzens and Roberts [to appear].

3. Seriation and Robinson form. In making decisions, it is often necessary to order the alternatives from least desirable to most desirable, least risky to most risky, etc. One way to order the alternatives begins with a judgement r_{ij} of the similarity of alternatives, with r_{ij} higher than r_{kl} if i and j are more similar than k and l. $R = (r_{ij})$ is a symmetric matrix, called the *proximity matrix*. This type of matrix has been used extensively in both the clustering and the multidimensional scaling literature. R is said to be in *strong Robinson form* if whenever $i \leqq j \leqq k \leqq l$ then $r_{jk} \geqq r_{il}$. This means that if j and k are between i and l then j and k are at least as similar as i and l. A goal of seriation (Hubert [1974]) is to find a permutation of the objects being sequenced so that if this permutation is applied simultaneously to both rows and columns of the proximity matrix R, the resulting matrix is in strong Robinson form. The ordering is thought of as the natural ordering determined by the similarity data. Kendall [1963], [1969a, b], [1971a, b, c] posed the questions: When does such a permutation exist, and how does one find it when it does exist?

Let δ be an arbitrary positive number, a threshold selected by the researcher, and define a graph G_δ from R as follows. The vertices of G are the numbers $1, 2, \cdots, n$, where n is the number of rows of R. There is an edge from i to j if and only if $r_{ij} \geqq \delta$. Let R' be obtained from R by replacing each diagonal element by ∞. Let G'_δ be defined from R' as G_δ was defined from R. We thus have the following theorem, which first appeared in Roberts [1979a] but is based heavily on the work of Hubert [1974]:

THEOREM 2. *A square symmetric matrix R is permutable to Robinson form if and only if the diagonal elements of R are each maximal in their row and $\{G'_\delta : \delta > 0\}$ is a homogeneous*[2] *family of indifference graphs.*

There are only a finite number of different graphs G'_δ, and therefore this condition is indeed testable.

The use of a threshold δ corresponds to dichotomizing the entries of the proximity matrix—some are over threshold, some are under threshold. In some instances measurement error may obscure an otherwise perfect model and a certain amount of insight is needed to determine what type of proximity matrix dichotomization may prove most useful. Hubert [1974] suggests that it may be valuable, in some cases, to trichotomize a proximity matrix as a way of lessening any inconsistencies due to measurement error. More precisely we might use two threshold levels, δ_1 and δ_2, $\delta_1 > \delta_2$, to be determined by the researcher. We shall study the pair of graphs G_{δ_1} and G_{δ_2}. We shall consider a problem analogous to the indifference graph characterization problem: If G and H are two graphs with the same vertex set V and δ_1 and δ_2 are two positive numbers, when does there exist a function $f: V \to \mathbb{R}$ such that (5) holds? The solution to this problem is clearly related to the trichotomization of similarity matrices in the same way the solution to the problem of characterizing indifference graphs was related to the dichotomization of such matrices.

4. Bisemiorders and Guttman scales. Another representation related to the representation (4) arises as follows. Suppose S is a set of individuals whose reactions or experiences are being studied and E is a set of reactions or experiences. Let aRb mean that individual a had reaction or experience b. In another interpretation E is a set of test questions and aRb means that individual a answers question b correctly. Then R defines a binary relation on $S \times E$ so that $R \subseteq S \times E$. Many experimenters

[2] A family $\{G'_\delta\}$ of indifference graphs on the same vertex set V is called *homogeneous* if there is a linear (simple) ordering $\leqq$ of the vertices in V so that for all δ, $x \leqq y \leqq z \leqq w$ and $\{x, w\} \in E(G'_\delta)$ implies $\{y, z\} \in E(G'_\delta)$.

have sought a way to order reactions (test questions) in such a way that if individual a experiences reaction b (answers question b correctly), he tended to experience all reactions (answer correctly all questions) coming before it in the order. In particular, they have sought a way to simultaneously order the individuals and reactions (questions) in such a way that individual a had reaction b (answered question b correctly) if and only if a followed b in the ordering. In terms of a representation, we can think of finding two real-valued functions s on S and e on E such that for all $a \in S$ and $b \in E$,

$$aRb \Leftrightarrow s(a) > e(b). \tag{11}$$

The two functions s and e satisfying (11) define what is called a *Guttman scale*, after Louis Guttman [1944]. In general, given a triple (S, E, R) with R a subset of $S \times E$, we ask when (S, E, R) possesses a Guttman scale. For a representation theorem, see Ducamp and Falmagne [1969].

Suppose we want to distinguish two types of reactions or replies, instead of just one. For example, an individual may show intense fear or just fear as a reaction; an individual may answer a question correctly with difficulty or correctly without difficulty. Let R and T be binary relations on A representing these two levels of reactions, or of correctness of replies, with $R \subseteq S \times E$ and $T \subseteq S \times E$. We now want conditions on R and T that will produce a representation, namely, functions $s: S \to \mathbb{R}$ and $e: E \to \mathbb{R}$, and positive numbers δ, η with $\delta > \eta$, such that for all $a \in S$, $b \in E$,

$$aRb \Leftrightarrow s(a) > e(b) + \delta \quad \text{and} \quad aTb \Leftrightarrow s(a) > e(b) + \eta. \tag{12}$$

The functions s and e provide a generalization of the Guttman scale to two relations, and the representation (12) is obviously closely related to our representation (4). Ducamp and Falmagne [1969] introduce the concept of *bisemiorder* (S, E, R, T) and show that (S, E, R, T) is representable in the form (12) if and only if it is a bisemiorder. The conditions for a bisemiorder do not solve the representation problem (4) because of the special nature of the relations R and T and because we have two functions rather than one.

5. Upper homogeneous representations. A question related to that asked in § 1 is the following: Suppose we are given a structure (A, P_1, P_2) where each P_i is an asymmetric binary relation on A which represents preference. We think of two levels of preference, strong or weak. When does there exist a function f and nonnegative threshold functions ϕ_1 and ϕ_2 defined on A such that for all $a, b \in A$:

$$aP_1b \Leftrightarrow f(a) > f(b) + \phi_1(b) \quad \text{and} \quad aP_2b \Leftrightarrow f(a) > f(b) + \phi_2(b)? \tag{13}$$

To answer this question we introduce the concept of *interval order*, a binary relation (A, P) satisfying Axioms S_1 and S_2 of the semiorder axioms. We note the following representation theorem due to Fishburn [1970].

THEOREM 3. *Suppose (A, P) is a binary relation on a countable set A. Then (A, P) is an interval order if and only if there exist real-valued functions f and ϕ defined on A, with $\phi(a) > 0$ for each $a \in A$, such that for all $a, b \in A$,*

$$aPb \Leftrightarrow f(a) > f(b) + \phi(b).$$

It is clear that for a representation (13) to exist, (A, P_1) and (A, P_2) must be interval orders. The following definitions are basically those of Krantz, Luce, Suppes and Tversky [to appear], but phrased along the lines previously used. Suppose P_1 and P_2 are asymmetric relations on A and f, ϕ_1, ϕ_2 are real-valued functions on A. Then

we call $\langle f, \phi_1, \phi_2\rangle$ an *upper homogeneous representation* of (A, P_1, P_2) if f, ϕ_1, ϕ_2 satisfy (13). We say P_1 and P_2 are *upper interval homogeneous* if for all $a, b, c, d \in A$, whenever aP_1b and cP_2d then either aP_2d or cP_1b. Notice that when $P_1 = P_2$ the condition reduces to S2 of the definition of an interval order. We now have the following theorem:

THEOREM 4 (Krantz et al. [to appear]). *Suppose (A, P_1) and (A, P_2) are interval orders and A is countable. Then (A, P_1, P_2) has an upper homogeneous representation if and only if P_1 and P_2 are upper interval homogeneous.*

The preceding leaves a number of questions yet unanswered. First, given an upper homogeneous representation $\langle f, \phi_1, \phi_2\rangle$, when does there exist a function g and constant functions $\phi_1' \equiv k_1$ and $\phi_2' \equiv k_2$ such that $\langle gf, \phi_1', \phi_2'\rangle$ is also an upper homogeneous representation of (A, P_1, P_2)? This question was posed by Krantz et al. in a preliminary version of [to appear, Chap. 15] and asks when it is possible to transform an upper homogeneous representation to one with constant thresholds. In the question posed at the end of § 1, we are asking the closely related question: Is there an upper homogeneous representation with ϕ_1 and ϕ_2 constant functions?

6. Precise formulation of the questions. We will ask two questions for each of the representations (4) and (5). Namely, we first ask if for a *fixed* $\delta_1 > \delta_2 > 0$ there is a function f satisfying (4) or (5). We next ask if for *some* $\delta_1 > \delta_2 > 0$ there is such a function. In the case of one threshold, as Corollary 1.1 pointed out, there is a representation (say satisfying (1)) with *fixed* $\delta > 0$ if and only if there is a representation with *some* $\delta > 0$. However, in the case of two thresholds, the questions are distinct. Consider the following example. Let $A = \{a, b, c, d\}$, $P_1 = \{(a, b)\}$ and $P_2 = \{(a, b), (a, d)\}$, $\delta_1 = 3$ and $\delta_2 = 1$. Then no real-valued function f exists such that (4) holds for $\delta_1 = 3$ and $\delta_2 = 1$ and all members of A. Since aP_1b we must have $f(a) > f(b) + 3$, but not aP_2c and not cP_2b implies $f(a) \leqq f(c) + 1$ and $f(c) \leqq f(b) + 1$. Thus $f(a) \leqq f(b) + 2$, a contradiction. For a function f to exist for (A, P_1) and (A, P_2) and satisfying (4), we must limit δ_1 and δ_2. We want $f(a) > f(b) + \delta_1$, $f(a) \leqq f(c) + \delta_2$, and $f(c) \leqq f(b) + \delta_2$. Therefore $f(b) + \delta_1 < f(a) \leqq f(b) + 2\delta_2$ and we must have $\delta_1 < 2\delta_2$. Since aP_2d and not aP_1d, $\delta_1 > \delta_2$. Therefore $\delta_2 < \delta_1 < 2\delta_2$. In other words the difference in thresholds must be fairly small to represent (A, P_1) and (A, P_2) simultaneously. One possible representation for the above example can be found with $\delta_1 = 1.5$ and $\delta_2 = .8$. Take $f(a) = 1.6$, $f(b) = 0$, $f(c) = .8$ and $f(d) = .7$.

7. Scott's method. In studying the representation (4) in the next section we will use a method developed by Scott [1964] to solve a large class of representation problems. He used this method to give an alternative proof of the representation of nontransitive indifference (semiorders), to find and prove conditions necessary and sufficient for the representation of ordered differences (Adams and Fagot [1956]), and to give an alternative solution to the problem of subjective probabilities (Kraft, Pratt, and Seidenberg [1959]). Since that time, the method has had a number of uses. In particular, Ducamp [1978] has used this same method to give an alternate proof of the representation theorem for bisemiorders (Ducamp and Falmagne [1969]).

Scott's method is based on the following ideas. Let L be a finite dimensional vector space over the reals. A subset $X \subseteq L$ is *symmetric* if $X = -X = \{-x \mid x \in X\}$. A subset $N \subseteq X$ is *realizable* in X if there is a linear functional ϕ on L such that for all $x \in X$,

$$x \in N \Leftrightarrow \phi(x) \geqq 0. \tag{14}$$

Recall that a linear functional is a real-valued, homogeneous, additive function defined on L. In case $N \neq X$, (14) says that there is a half space H of L separating the sets

N and $X-N$ so that $X \cap H = H \cap N$. If $N = X$ then the trivial linear functional $\phi \equiv 0$ shows that N is realizable. We let M denote $X-N$. In the case that S is a finite set and $L = L(S)$ is the vector space of all real-valued functions defined on S (the usual S-dimensional vector space), we call a vector (function) in L *rational* if all its coordinates (values) are rational numbers. A *rational subset of* L is a subset of L consisting entirely of rational vectors. With the preceding definitions we can now state the main theorem of Scott's method.

THEOREM 5 (Scott [1964]). *Let L be a finite dimensional vector space, and X a finite, rational, symmetric subset of L. For a subset N of X to be realizable in X it is necessary and sufficient that the conditions*

$$x \in N \vee -x \in N$$

and

$$\sum_{i=0}^{n-1} x_i = 0 \Rightarrow -x_0 \in N$$

hold for all $x \in X$ and all sequences $x_0, x_1, \cdots, x_{n-1} \in X$ where $x_i \in N$ for all $i < n$, $i > 0$, and $n > 0$.

We will use a particular case of Theorem 5 which we will now state as a corollary.

COROLLARY 5.1. *Let L be a finite dimensional vector space. Let M, N, X be such that X is a finite, rational, symmetric subset of L, (M, N) a partition of X and $N = -M$. Then the following are equivalent*:

(i) *There exists a linear functional h on L such that for all $x \in X$, $x \in M$ if and only if $h(x) > 0$.*

(ii) *There exist no sequences $x_1, x_2, \cdots, x_n$ in M such that $\sum_{i=1}^{n} x_i = 0$.*

8. Double semiorders. If A is a set and P_1 and P_2 are binary relations in A, let us say that (A, P_1, P_2) is *two-threshold representable*, or *representable* for short, if there exists a real-valued function f on A and two positive real numbers δ_1 and δ_2, $\delta_1 > \delta_2$, so that for all $a, b \in A$, (4) holds. We call f a *two-threshold representation* for (A, P_1, P_2) relative to δ_1 and δ_2. Note that if $\delta_1 = \delta_2$, then $P_1 = P_2$ and the Scott–Suppes theorem (Theorem 1) is a representation theorem for (A, P_1, P_2). That is why we consider only the case $\delta_1 > \delta_2$.

By the Scott–Suppes theorem, if (A, P_1, P_2) is representable, P_1 and P_2 must both be semiorders. It is also clear that $\delta_1 > \delta_2$ implies that $P_1 \subseteq P_2$. In this case we say P_1 and P_2 form a *nested pair of semiorders.*

We now need to look at the relationship of δ_1 to δ_2 if (A, P_1, P_2) is to be representable with $\delta_1 > \delta_2 > 0$. At this stage it becomes convenient to adopt some notation which we will use throughout this section. If (A, R) is a binary relation, let $\bar{R}^* = \{(y, x) \mid (x, y) \in A \times A \text{ and } (x, y) \notin R\}$. A *cycle* C in (A, P_1, P_2) is a sequence $R_1, R_2, \cdots, R_n$ such that $R_i \in \{P_1, P_2, \bar{P}_1^*, \bar{P}_2^*\}$ for each i, and there exist $a_1, a_2, \cdots, a_n \in A$ such that $a_1 R_1 a_2 R_2 \cdots R_{n-1} a_n R_n a_1$. Consider two nested semiorders, $P_1 \subseteq P_2$, each defined on set A. For any cycle C in (A, P_1, P_2), suppose $m_1 = m_1(C)$ is the number of P_1 in the cycle C, $m_2 = m_2(C)$ is the number of P_2 in the cycle C, $n_1 = n_1(C)$ is the number of $\bar{P}_1^*$ in the cycle C, and $n_2 = n_2(C)$ is the number of $\bar{P}_2^*$ in the cycle C. Let

$$m_C(P_1, P_2) = \begin{cases} \dfrac{m_2 - n_2}{n_1 - m_1} & \text{if } C \text{ is a cycle such that } n_1 > m_1 \text{ and } m_2 \geqq n_2, \\ 0 & \text{otherwise,} \end{cases}$$

$$s_C(P_1, P_2) = \begin{cases} \dfrac{n_2 - m_2}{m_1 - n_1} & \text{if } C \text{ is a cycle such that } m_1 > n_1, \\ \infty & \text{otherwise.} \end{cases}$$

Define $m(P_1, P_2)$ to be $\sup_C m_C(P_1, P_2)$ and $s(P_1, P_2)$ to be $\inf_C s_C(P_1, P_2)$ if $s_C(P_1, P_2) < \infty$ for some C and to be ∞ otherwise. Note that in principle $s(P_1, P_2)$ could be $-\infty$. However, we shall show that representability implies $s(P_1, P_2) \geqq 0$.

LEMMA 1. *If (A, P_1) and (A, P_2) are nested semiorders with $P_1 \subsetneqq P_2$, then $m(P_1, P_2) \geqq 1$.*

Proof. Since $P_1 \subsetneqq P_2$, there exist $a, b \in A$ such that aP_2b and $\sim aP_1b$. Therefore we always have the cycle $C' = aP_2b\bar{P}_1^*a$, and $m_{C'}(P_1, P_2) = \frac{1}{1} = 1$. Therefore $m(P_1, P_2) \geqq m_{C'}(P_1, P_2) = 1$. Q.E.D.

Let us look at an example that shows that *not* all nested semiorder systems are representable. Let $A = \{a, b, c, d, e\}$, $P_1 = \{(b, a)\}$, and $P_2 = \{(b, a), (b, c), (b, d), (c, d), (e, d)\}$. This example is illustrated with the multidigraph of Fig. 1, where two arcs from x to y indicate xP_1y and xP_2y and one arc indicates xP_2y and not xP_1y. Suppose a representation f exists for (A, P_1, P_2) with some $\delta_1 > \delta_2 > 0$. Now bP_2cP_2d implies $f(b) > f(c) + \delta_2 > f(d) + 2\delta_2$. But $\sim bP_1d$ implies $f(b) \leqq f(d) + \delta_1$. Therefore $f(d) + 2\delta_2 < f(b) \leqq f(d) + \delta_1$, so $2\delta_2 < \delta_1$. Now $\sim eP_2a$ and $\sim bP_2e$ implies $f(b) \leqq f(e) + \delta_2 \leqq f(a) + 2\delta_2$. Also $f(b) > f(a) + \delta_1$ since bP_1a. Therefore $f(a) + \delta_1 < f(b) \leqq f(a) + 2\delta_2$ so $\delta_1 < 2\delta_2$ and we have $\delta_1 < 2\delta_2 < \delta_1$, a contradiction. Yet both P_1 and P_2 are semiorders, and $P_1 \subseteq P_2$.

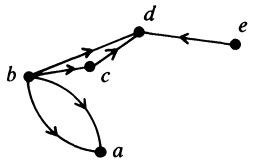


FIG. 1

In this example, the cycle C_1 in (A, P_1, P_2) defined by $C_1 = bP_2cP_2d\bar{P}_1^*b$ is such that $m_{C_1}(P_1, P_2) = \frac{2}{1} = 2$. Also the cycle $C_2 = bP_1a\bar{P}_2^*e\bar{P}_2^*b$ is such that $s_{C_2}(P_1, P_2) = \frac{2}{1} = 2$. Nonrepresentability now follows from the following lemma.

LEMMA 2. *Suppose (A, P_1, P_2) is representable by a function f with $\delta_1 > \delta_2 > 0$ and P_1 and P_2 are semiorders such that $P_1 \subseteq P_2$. Then for all cycles C and C' in (A, P_1, P_2) such that $s_C(P_1, P_2) < \infty$,*

$$s_C(P_1, P_2) \geqq \frac{\varepsilon_1}{\delta_2} > \frac{\delta_1}{\delta_2} \geqq \frac{\varepsilon_1'}{\delta_2} \geqq m_{C'}(P_1, P_2),$$

where $\varepsilon_1 = \min\{f(a) - f(b): aP_1b\}$ and $\varepsilon_1' = \max\{f(a) - f(b): \sim aP_1b\}$.

Proof. Since $s_C(P_1, P_2) < \infty$ there are $a, b \in A$ such that aP_1b. Thus ε_1 is defined. Also, clearly, $\varepsilon_1 > \delta_1$ so $\varepsilon_1/\delta_2 > \delta_1/\delta_2$. Similarly, $\sim aP_1a$ implies ε_1' is defined and $\varepsilon_1' \geqq 0$. Clearly $\delta_1 \geqq \varepsilon_1'$, so $\delta_1/\delta_2 \geqq \varepsilon_1'/\delta_2$.

Let C be the cycle $a_1R_1a_2R_2 \cdots R_{k-1}a_kR_ka_1$ such that $R_i \in \{P_1, P_2, \bar{P}_1^*, \bar{P}_2^*\}$ for each $i = 1, 2, \cdots, k$, $a_j \in A$ for each $j = 1, 2, \cdots, k$ and $s_C(P_1, P_2) < \infty$. Let C' be the cycle $b_1R_1'b_2R_2' \cdots R_{l-1}'b_lR_l'b_1$ where $R_i' \in \{P_1, P_2, \bar{P}_1^*, \bar{P}_2^*\}$, for each $i = 1, 2, \cdots, l$, and $b_j \in A$ for each $j = 1, 2, \cdots, l$. For all $x, y \in A$, $xP_1y \Rightarrow f(x) \geqq f(y) + \varepsilon_1$, $x\bar{P}_1^*y \Rightarrow \sim yP_1x \Rightarrow f(y) \leqq f(x) + \delta_1 \Rightarrow f(x) \geqq f(y) - \delta_1 \Rightarrow f(x) \geqq f(y) - \varepsilon_1$, $xP_2y \Rightarrow f(x) > f(y) + \delta_2$, and $x\bar{P}_2^*y \Rightarrow \sim yP_2x \Rightarrow f(y) \leqq f(x) + \delta_2 \Rightarrow f(x) \geqq f(y) - \delta_2$. Thus if $R_i = P_1$

then $f(x) \geqq f(y)+\varepsilon_1$ and if $R_i = \bar{P}_1^*$ then $f(x) \geqq f(y)-\varepsilon_1$. Also if $R_i = P_2$ then $f(x) > f(y)+\delta_2$ and if $R_i = \bar{P}_2^*$ then $f(x) \geqq f(y)-\delta_2$. Thus from the cycle $C = a_1 R_1 a_2 R_2 \cdots a_{k-1} R_{k-1} a_k R_k a_1$ we have $f(a_1) \geqq f(a_1)+m_1\varepsilon_1+m_2\delta_2-n_1\varepsilon_1-n_2\delta_2$, where m_i is the number of P_i in the cycle C and n_i is the number of $\bar{P}_i^*$ in the cycle C, $i = 1, 2$. Therefore $m_1\varepsilon_1+m_2\delta_2-n_1\varepsilon_1-n_2\delta_2 \leqq 0$. Since $s_C(P_1, P_2) < \infty$, $m_1 > n_1 \geqq 0$. Thus $\varepsilon_1/\delta_2 \leqq (n_2-m_2)/(m_1-n_1)$. Hence $\varepsilon_1/\delta_2 \leqq s_C(P_1, P_2)$, proving the first inequality.

To prove the fourth (last) inequality, note that by a similar argument from the cycle $C' = b_1 R_1' b_2 R_2' \cdots R_{l-1}' b_l R_l' b_1$ we have $f(b_1) \geqq f(b_1)+m_1'\varepsilon_1'+m_2'\delta_2-n_1'\varepsilon_1'-n_2'\delta_2$, where m_i' = number of P_i in the cycle C', $i = 1, 2$, and n_i' = number of $\bar{P}_i^*$ in the cycle C', $i = 1, 2$. Now suppose $m_{C'}(P_1, P_2) = 0$. Then certainly $m_{C'}(P_1, P_2) \leqq \varepsilon_1'/\delta_2$. If $m_{C'}(P_1, P_2) \neq 0$ then $n_1' > m_1' \geqq 0$ and $m_2' \geqq n_2' \geqq 0$. Thus $f(b_1) \geqq f(b_1)+m_1'\varepsilon_1'+m_2'\delta_2-n_1'\varepsilon_1'-n_2'\delta_2$ implies that $m_1'\varepsilon_1'+m_2'\delta_2-n_1'\varepsilon_1'-n_2'\delta_2 \leqq 0$. Hence, $(m_2'-n_2')/(n_1'-m_1') \leqq \varepsilon_1'/\delta_2$, since $n_1'-m_1' > 0$. Therefore $m_{C'}(P_1, P_2) \leqq \varepsilon_1'/\delta_2$ and the last inequality is proved. Q.E.D.

Note that it follows from Lemma 2 that if (A, P_1, P_2) is representable and $s_C(P_1, P_2) < \infty$, then $s_C(P_1, P_2) \geqq 0$. Thus, if $s_C(P_1, P_2) < \infty$ for some C, $s(P_1, P_2) \geqq 0$.

LEMMA 3. *If (A, P_1, P_2) is representable, we have $m(P_1, P_2) < s(P_1, P_2)$.*

Proof. The result follows from Lemma 2 if $s_C(P_1, P_2) < \infty$ for some cycle C. If the latter is false, then $s(P_1, P_2) = \infty$ and $m(P_1, P_2) < s(P_1, P_2)$ is immediate because by Lemma 2, $m(P_1, P_2) \leqq \delta_1/\delta_2$. Q.E.D.

Although the condition that $s(P_1, P_2)$ be greater than $m(P_1, P_2)$ seems to be a difficult condition to verify, there are situations where the condition is useful, as in Fig. 1 above. To generalize this example, define a *P_2-path* in the multidigraph depicting a system (A, P_1, P_2) as a cycle $C = a_0 P_2 a_1 P_2 \cdots P_2 a_k \bar{P}_1^* a_0$. P_2-paths are readily identified and $m_C(P_1, P_2) = k/1 = k$ for a P_2-path with $k+1$ elements of A. Therefore $m(P_1, P_2) \geqq$ length of longest P_2-path in (A, P_1, P_2).

Define an *(I_2-P_1)-cycle* as a cycle $C = a_0 \bar{P}_2^* a_1 \bar{P}_2^* \cdots a_{l-1} \bar{P}_2^* a_l P_1 a_0$. For such a cycle $s_C(P_1, P_2) = l/1 = l$, and so $s(P_1, P_2) \leqq l$. Therefore $s(P_1, P_2) \leqq$ length of shortest (I_2-P_2)-cycle.

Looking back at Fig. 1, we found $m_{C_1}(P_1, P_2) = 2$ from the P_2-path $C_1 = bP_2cP_2d\bar{P}_1^*b$, but $s_{C_2}(P_1, P_2) = 2$ from the (I_2-P_1)-cycle $C_2 = a\bar{P}_2^*e\bar{P}_2^*bP_1a$. Thus $m_{C_1}(P_1, P_2) \not< s_{C_2}(P_1, P_2)$, so $m(P_1, P_2) \not< s(P_1, P_2)$, hence no representation exists for this (A, P_1, P_2) by Lemma 3. In general we can use Lemma 3 to prove nonrepresentability by taking long P_2-paths and short (I_2-P_1)-cycles.

The next lemma gives conditions for (A, P_1, P_2) to be representable in the trivial case that $P_1 = P_2$.

LEMMA 4. *If $P_1 = P_2$, A is a finite set and (A, P_2) is a semiorder, then (A, P_1, P_2) is representable.*

Proof. If P_2 is a semiorder then by the Scott–Suppes theorem (Theorem 1), given $\delta_2 > 0$, there exists a function $f: A \to R$ such that for all $a, b \in A$,

$$aP_2b \Leftrightarrow f(a) > f(b)+\delta_2.$$

Let $\varepsilon = \min\{f(a)-f(b): aP_2b\}$ and $\delta_1 = \delta_2+(\varepsilon-\delta_2)/2 = (\delta_2+\varepsilon)/2$. Then it is easy to show that f is a two-threshold representation relative to δ_1 and δ_2. Q.E.D.

Now we will use the techniques of Scott's method to deduce the remaining conditions both necessary and sufficient for a system (A, P_1, P_2) to be representable where $P_1 \neq P_2$. We already know that P_1 and P_2 must be nested semiorders, and $m(P_1, P_2) < s(P_1, P_2)$.

Let (A, P_1, P_2) be a system of two semiorders P_1, P_2 such that $P_1 \subsetneq P_2$, A is finite, and $m(P_1, P_2) < s(P_1, P_2)$. Let e be an element not in the set A and let $S = A \cup \{e\}$.

Each element $x \in S$ determines a vector in $L = L(S)$, namely the characteristic function of $\{x\}$. We identify each $x \in S$ with its corresponding vector, so $S \subseteq L$. Now S becomes a linear basis in $L(S)$. Choose a rational number δ such that $m(P_1, P_2) < \delta < s(P_1, P_2)$. Since $m(P_1, P_2) \geqq 1$ by Lemma 1, note that $\delta > 1$. Let:

$$M_1 = \{a - (b + \delta e) \mid a, b \in A, aP_1b\},$$

$$M_2 = \{a - (b + e) \mid a, b \in A, aP_2b\},$$

$$\bar{M}_1 = \{(b + \delta e) - a \mid a, b \in A, \sim aP_1b\},$$

$$\bar{M}_2 = \{(b + e) - a \mid a, b \in A, \sim aP_2b\}.$$

Define $M = M_1 \cup M_2 \cup \bar{M}_1 \cup \bar{M}_2$ and $N = -M$. Define $X = M \cup N$. Now $X = M \cup N$ and X is a finite, rational, symmetric subset of $L(S)$ and (M, N) is a partition of X. We can now use Scott's method, Corollary 5.1, to find conditions for the existence of a linear functional h on L such that for all $x \in X$, $x \in M$ if and only if $h(x) > 0$. Such a function h exists if and only if there is no sequence $x_1, x_2, \cdots, x_n \in M$ such that $\sum_{i=1}^{n} x_i = 0$. We shall derive conditions to ensure this. Since h is a linear functional,

$$\begin{aligned} h(a) > h(b) + \delta(h(e)) &\Leftrightarrow h(a - b - \delta e) > 0 \\ &\Leftrightarrow a - b - \delta e \in M \\ &\Leftrightarrow aP_1b \end{aligned}$$

and

$$\begin{aligned} h(a) > h(b) + h(e) &\Leftrightarrow h(a - b - e) > 0 \\ &\Leftrightarrow a - b - e \in M \\ &\Leftrightarrow aP_2b. \end{aligned}$$

Since $P_1 \subsetneqq P_2$, there exist $a_0, b_0 \in A$ such that $a_0P_2b_0$ and $\sim a_0P_1b_0$. Therefore we have $h(a_0) > h(b_0) + h(e)$ and $h(a_0) \leqq h(b_0) + \delta(h(e))$. Thus $h(e) < \delta(h(e))$ and $0 < h(e)(\delta - 1)$. Since $\delta > 1$, $h(e) > 0$. For any $a \in A$, let $f(a) = h(a)/h(e)$. Now

$$aP_1b \Leftrightarrow \frac{h(a)}{h(e)} > \frac{h(b)}{h(e)} + \delta \Leftrightarrow f(a) > f(b) + \delta$$

and

$$aP_2b \Leftrightarrow \frac{h(a)}{h(e)} > \frac{h(b)}{h(e)} + 1 \Leftrightarrow f(a) > f(b) + 1.$$

Thus the existence of h implies that there exists a function $f: A \to \mathbb{R}$ and $\delta_1 = \delta$, $\delta_2 = 1$ such that (4) is satisfied. Therefore (A, P_1, P_2) is representable. Conversely, if there is f, we can define h by working backwards. Thus we have proved the following lemma.

LEMMA 5. *Suppose P_1 and P_2 are a nested pair of semiorders on a finite set A, $P_1 \subsetneqq P_2$, and $m(P_1, P_2) < \delta < s(P_1, P_2)$, δ rational. Then (A, P_1, P_2) is representable with $\delta_1 = \delta$ and $\delta_2 = 1$ if and only if there is no sequence $x_1, x_2, \cdots, x_n \in M$ such that $\sum_{i=1}^{n} x_i = 0$.*

To discover conditions necessary and sufficient for representability, we suppose that there is a sequence $x_1, x_2, \cdots, x_n$ in M such that $\sum_{i=1}^{n} x_i = 0$. Now any element of M is of the form $x - y + \varepsilon$, where $x, y \in A$, $\varepsilon = \pm e$ or $\pm\delta e$. Let $x_{i_1} = a_0 - a_1 + \varepsilon_1$, with $a_0, a_1 \in A$, $\varepsilon_1 = \pm e$ or $\pm\delta e$. Since the algebraic sum of the component a_1 in $\sum x_i$ must be 0 (i.e., $\sum x_i$ has 0 as the coefficient of a_1 in the expansion), there must be

$x_{i_2} = a_1 - a_2 + \varepsilon_2$ where $a_2 \in A$ and $\varepsilon_2 = \pm e$ or $\pm\delta e$. If $a_2 \neq a_0$, we can find $x_{i_3} = a_2 - a_3 + \varepsilon_3$ where $a_3 \in A$ and $\varepsilon_3 = \pm e$ or $\pm\delta e$. Continuing, we get a subsequence (possibly the whole sequence) $\{x_{i_j}\}_{1 \leqq j \leqq m}$ such that for $1 \leqq j \leqq m$, $x_{i_j} = a_{j-1} - a_j + \varepsilon_j$ and $\varepsilon_j = \pm e$ or $\pm\delta e$, where a_m means a_0. The subsequence corresponds to a "cycle" $a_0, a_1, \cdots, a_{m-1}, a_m = a_0$. We may suppose $\sum_{j=1}^{m} x_{i_j} \leqq 0$. For if not, then $\sum_{j=1}^{m} x_{i_j} > 0$ and $m \neq n$. We can repeat the process choosing a distinct $x_{i_{m+1}}$ ($x_{i_{m+1}} \neq x_{i_j}$ for all $j = 1, 2, \cdots, m$), generating a new "cycle". Specifically, choose $x_{i_{m+1}} \neq x_{i_j}$, $j \leqq m$. Let $x_{i_{m+1}} = a_m - a_{m+1} + \varepsilon_{m+1}$, where $\varepsilon_{m+1} = \pm e$ or $\pm\delta e$. Since the algebraic sum of the component a_m in $\sum x_i$ must be 0 there must be $x_{i_{m+2}} \neq x_{i_j}$ for all $j \leqq m$ of the form $x_{i_{m+2}} = a_{m+1} - a_{m+2} + \varepsilon_{m+2}$, where $\varepsilon_{m+2} = \pm e$ or $\pm\delta e$. Since the sum, $\sum x_i$, is 0, for each $x_{i_{m+2}}$ there also must exist an $x_{i_{m+3}} = a_{m+2} - a_{m+3} + \varepsilon_{m+3}$, and so on. We thus find a new subsequence $x_{i_{m+1}}, x_{i_{m+2}}, \cdots, x_{i_{m+m'}}$, with properties analogous to the subsequence $x_{i_1}, x_{i_2}, \cdots, x_{i_m}$. If $\sum_{k=m+1}^{m+m'} x_{i_k} \leqq 0$, we use the new subsequence. If $\sum_{k=m+1}^{m+m'} x_{i_k} > 0$, we find a third subsequence $x_{i_{m+m'+1}}, x_{i_{m+m'+2}}, \cdots, x_{i_{m+m'+m''}}$. Since all subsequences have disjoint elements, and since M is finite and $\sum x_i = 0$, we must eventually find a subsequence whose sum is $\leqq 0$. We denote this subsequence by $\{x_{i_j}\}_{j=1}^{m}$ and let $x_{i_j} = a_{j-1} - a_j + \varepsilon_j$.

Note that for $j \geqq 1$, $\varepsilon_j = -e$ implies $a_{j-1} P_2 a_j$; $\varepsilon_j = -\delta e$ implies $a_{j-1} P_1 a_j$; $\varepsilon_j = e$ implies $\sim a_j P_2 a_{j-1}$, equivalently $a_{j-1} \bar{P}_2^* a_j$, and $\varepsilon_j = \delta e$ implies $\sim a_j P_1 a_{j-1}$, equivalently $a_{j-1} \bar{P}_1^* a_j$. In other words, the subsequence corresponds to a cycle $C = a_0 R_1 a_1 R_2 \cdots R_{m-1} a_{m-1} R_m a_0$ where $R_i \in \{P_1, P_2, \bar{P}_1^*, \bar{P}_2^*\}$ for each i. The cycle C does not contain only P_1 and P_2 since $P_1 \subsetneqq P_2$ and P_2 is transitive and irreflexive. Let m_i denote the number of P_i in the cycle and n_i denote the number of $\bar{P}_i^*$ in the cycle. Now $\sum_{j=1}^{m} x_{i_j} = -m_1 \delta e + n_1 \delta e - m_2 e + n_2 e \leqq 0$, so $(n_1 - m_1)\delta e \leqq (m_2 - n_2)e$. If $n_1 > m_1$, then $\delta \leqq (m_2 - n_2)/(n_1 - m_1) \leqq m_C(P_1, P_2)$, contradicting $m(P_1, P_2) < \delta$. If $m_1 > n_1$, then $s_C(P_1, P_2) < \infty$ for cycle C corresponding to the subsequence. Now $\delta \geqq (m_2 - n_2)/(n_1 - m_1) = (n_2 - m_2)/(m_1 - n_1) \geqq s(P_1, P_2)$, contradicting $\delta < s(P_1, P_2)$. Therefore $m_1 = n_1$. It follows that $m_2 \geqq n_2$. Consider the cycle $R_1 R_2 \cdots R_m$ (the a_i's will be suppressed when not needed). Assume first that this cycle contains only P_1 and $\bar{P}_1^*$ or only P_2 and $\bar{P}_2^*$, abbreviated P and $\bar{P}^*$. Since aPa is a contradiction and since $aPb\bar{P}^*a$ is a contradiction (it says aPb and $\sim aPb$), there must be at least two P's in the cycle. But if there are two P's adjacent in the cycle, then there are two P's followed or preceded by a $\bar{P}^*$. Thus we have $xPyPz\bar{P}^*w$ or $x\bar{P}^*yPzPw$, both of which imply xPw by the semiorder axioms. By replacing $xPyPz\bar{P}^*w$ or $x\bar{P}^*yPzPw$ by xPw we reduce the cycle $R_1 R_2 \cdots R_m$ to a shorter one that satisfies $m_1 = n_1$ and $m_2 \geqq n_2$. Thus we can continuously reduce the cycle to one with only one P and one $\bar{P}^*$, already impossible, or to one with no two adjacent P's. If the latter, a part of the cycle looks like $P\bar{P}^*P$. But the semiorder axioms also yield $xPy\bar{P}^*zPw \Rightarrow xPw$, thereby reducing the cycle to a shorter one in which $m_1 = n_1$ and $m_2 \geqq n_2$ still holds. Thus the cycle can be systematically reduced to xPx, which we already know is impossible. Therefore we conclude that the cycle cannot contain only P_1 and $\bar{P}_1^*$ or only P_2 and $\bar{P}_2^*$. It follows that $m_1 = n_1 \neq 0$ and $m_2 \neq 0$.

Since $a\bar{P}_2^*a$ for all $a \in A$, if $m_2 > n_2$ we can add $\bar{P}_2^*$ to the cycle at any point and continue to have a cycle satisfying $m_1 = n_1 \neq 0$, $m_2 \geqq n_2$, $m_2 \neq 0$. Therefore we may as well assume that $m_1 = n_1$ and $m_2 = n_2$, and each is nonzero.

To sum up, if there is no representation, then by Lemma 5 there is a sequence $x_1, x_2, \cdots, x_n \in M$ such that $\sum x_i = 0$. We have shown that $\sum x_i = 0$ implies that there is a cycle $R_1 R_2 \cdots R_m$ such that $n_1 = m_1 \neq 0$, $n_2 = m_2 \neq 0$. Let us call such a cycle *balanced*. Thus we have shown for the case $P_1 \neq P_2$ that if (A, P_1, P_2) is a system of two semiorders P_1 and P_2 with $P_1 \subseteq P_2$ and $m(P_1, P_2) < s(P_1, P_2)$, and if there is no

representation, then there is a balanced cycle. The same result follows for the case $P_1 = P_2$ by Lemma 4. Conversely, if there is a representation f with δ and 1, we already know that P_1 and P_2 are semiorders, $P_1 \subseteq P_2$ and $m(P_1, P_2) < s(P_1, P_2)$. Also, there can be no balanced cycle. For suppose $a_1 R_1 a_2 R_2 \cdots a_k R_k a_1$ is a balanced cycle. Then $m_1 = n_1 \neq 0$ and $m_2 = n_2 \neq 0$ and, reasoning as in the proof of Lemma 2, $f(a_1) \geqq f(a_1) + m_1\delta + m_2 - n_1\delta - n_2$. Moreover, since $m_1 \neq 0$, there is at least one P_1 in the cycle, and so in fact the reasoning as in the proof of Lemma 2 gives us $f(a_1) > f(a_1) + m_1\delta + m_2 - n_1\delta - n_2$. Thus $f(a_1) > f(a_1) + (m_1 - n_1)\delta + (m_2 - n_2) = f(a_1)$, or $f(a_1) > f(a_1)$, a contradiction.

To state a representation theorem, let us say that (A, P_1, P_2) is a *balanced double semiorder system* if (A, P_1) and (A, P_2) are semiorders, $P_1 \subseteq P_2$, and (A, P_1, P_2) has no balanced cycle. We now have the following theorem.

THEOREM 6. *If A is a finite nonempty set, the system (A, P_1, P_2) is representable if and only if (A, P_1, P_2) is a balanced double semiorder system and $m(P_1, P_2) < s(P_1, P_2)$.*

To summarize the results so far, recall that initially two questions were asked. The first asked, given δ_1 and δ_2 and a system (A, P_1, P_2), when does there exist a function $f: A \to \mathbb{R}$ such that (4) holds? The second question modified the first to allow δ_1 and δ_2 to be chosen and asked: When is a system (A, P_1, P_2) representable? We have explicitly answered the second question in Theorem 6. The answer to the first question comes out of the proof of Theorem 6.

THEOREM 7. *Suppose A is a finite nonempty set and let δ_1 and δ_2 be given positive constants with $\delta_1 > \delta_2$. A system (A, P_1, P_2) is representable with δ_1, δ_2 if and only if (A, P_1, P_2) is a balanced double semiorder system and*

$$s(P_1, P_2) > \frac{\delta_1}{\delta_2} > m(P_1, P_2).$$

Proof. Representability with δ_1, δ_2 implies balanced double semiorder system by Theorem 6. That $s(P_1, P_2) > \delta_1/\delta_2 \geqq m(P_1, P_2)$ follows from Lemma 2. To see that the second inequality must be strict, note that we may modify f so that $f(a) - f(b)$ never equals δ_1. For suppose $f(x) - f(y) = \delta_1$. Then we may let $g(z)$ be $f(z)$ if $f(z) < f(x)$ and $g(z)$ be $f(z) - \varepsilon$ otherwise. If ε is picked small enough, then g is a representation and, moreover, $g(x) - g(y) < \delta_1$, and $g(a) - g(b) = \delta_1 \Rightarrow f(a) - f(b) = \delta_1$. We then make such modifications successively until we obtain a representation h with $h(a) - h(b)$ never equal to δ_1. Now it follows that $\delta_1 > \varepsilon_1'$ and so

$$\frac{\delta_1}{\delta_2} > \frac{\varepsilon_1'}{\delta_2} \geqq m(P_1, P_2).$$

To prove the converse, note that if $P_1 = P_2$ then representability follows from Lemma 4. If $P_1 \subsetneqq P_2$, the proof of Theorem 6 shows that if $s(P_1, P_2) > \delta > m(P_1, P_2)$, δ is rational, and (A, P_1, P_2) is not representable with $\delta_1 = \delta$ and $\delta_2 = 1$, then (A, P_1, P_2) is not a balanced double semiorder system. Thus if (A, P_1, P_2) is a balanced double semiorder system and $s(P_1, P_2) > \delta > m(P_1, P_2)$, and δ is rational, (A, P_1, P_2) is representable with $\delta_1 = \delta$ and $\delta_2 = 1$. Since this is true for rational δ, it clearly is true for all real δ. Now suppose (A, P_1, P_2) is a balanced double semiorder system and $s(P_1, P_2) > \delta_1/\delta_2 > m(P_1, P_2)$. Then (A, P_1, P_2) is representable with δ_1/δ_2 and 1, which by multiplying the representing function f by δ_2 gives us a representation with δ_1 and δ_2. Q.E.D.

COROLLARY 7.1. *A balanced double semiorder system (A, P_1, P_2) with $P_1 \neq P_2$ is representable for all $\delta_1 > \delta_2 > 0$ if and only if $m(P_1, P_2) = 1$ and $s(P_1, P_2) = \infty$.*

Proof. Theorem 7, Lemma 1 and the observation that $\delta_1 > \delta_2 > 0$ implies $\delta_1/\delta_2 > 1$. Q.E.D.

COROLLARY 7.2. *A balanced double semiorder system* (A, P_1, P_2) *is representable for all* $\delta_1 > \delta_2 > 0$ *if and only if* $m(P_1, P_2) \leqq 1$ *and* $s(P_1, P_2) = \infty$.

Proof. If $P_1 \subsetneqq P_2$, the result follows from Corollary 7.1. If $P_1 = P_2$, the result follows from Theorem 7 and the observation that $\delta_1 > \delta_2 > 0$ implies $\delta_1/\delta_2 > 1$. Q.E.D.

Now that the main results have been stated, and their proofs precede their statements in the discovery mode Scott suggests, let us examine the conditions further. In particular, suppose (A, P_1, P_2) is a balanced double semiorder system and consider a cycle $R_1R_2 \cdots R_m$. This is a cycle such that there exist $a_1, a_2, \cdots, a_m$ such that $a_1R_1a_2R_2 \cdots a_mR_ma_1$. Now if among $R_1, R_2, \cdots, R_{m-1}$ we have m_1' P_1's, m_2' P_2's, n_1' $\bar{P}_1^*$'s and n_2' $\bar{P}_2^*$'s with $m_2' = n_2' > 0$ and $m_1' = n_1' + 1$, then since $R_1R_2 \cdots R_m$ cannot be a balanced cycle, R_m cannot be $\bar{P}_1^*$. That is, $\sim a_m \bar{P}_1^* a_1$ or, equivalently, $a_1P_1a_m$. Thus by changing the order of the terms in a cycle, it is easy to see that the nonexistence of a balanced cycle is equivalent to the following *strong double semiorder condition*: If $a_1R_1a_2R_2 \cdots R_{m-1}a_m$ and $m_2' = n_2' > 0$ and $m_1' = n_1' + 1$, then $a_1P_1a_m$. Note that if $P_1 = P_2$, then the strong double semiorder condition is equivalent to the semiorder axioms.

In one of its simplest forms, the strong double semiorder condition states that for all $x, y, z, w \in A$, $xP_2y\bar{P}_2^*zP_1w \Rightarrow xP_1w$. This is equivalent to the upper interval homogeneous condition (§ 5). Therefore if (A, P_1, P_2) contains no balanced cycle, then (A, P_1, P_2) is upper interval homogeneous.

Suppose (A, P) is an asymmetric relation and I is the *symmetric complement* of P ($xIy \Leftrightarrow \sim xPy \wedge \sim yPx$) and (A, W) is a weak order. Then W is *compatible* with P if for all $x, y, z \in A$:

(i) $xPy \Rightarrow xWy$

and

(ii) $xWyWz \wedge xIz \Rightarrow xIy \wedge yIz$.

Roberts [1971b] has shown that if (A, P) is a semiorder, then there is an essentially unique weak order W on A compatible with P, and if an asymmetric relation (A, P) has a compatible weak order W, then (A, P) is a semiorder. A family of semiorders $\{(A, P_i)\}_i$ is *homogeneous* if there is a single weak order (A, W) compatible with (A, P_i) for each i. This concept was introduced by Roberts [1971a] in studying probabilistic consistency. If (A, P_1, P_2) is representable by a function f, then $\{(A, P_i)\}_i$ is a homogeneous family of semiorders. For define W on A by aWb iff $f(a) \geqq f(b)$. Then W is a weak order on A compatible with each P_i.

This second notion of homogeneity implies the first.

THEOREM 8. *If* $\{(A, P_i)\}_{i=1,2}$ *is a pair of homogeneous semiorders, then* (A, P_1) *and* (A, P_2) *are upper interval homogeneous.*

Proof. Suppose $\{(A, P_i)\}_{i=1,2}$ is a pair of homogeneous semiorders. Let I_i be the symmetric complement of P_i, $i = 1, 2$. Suppose aP_1b and cP_2d and $\sim aP_2d$ and $\sim cP_1b$. Let W be the weak order compatible with both P_1 and P_2. Then either cWb or bWc.

Case 1. bWc. Then (i) of compatibility implies $aWbWcWd$, and by condition (ii) of compatibility aI_2d would imply cI_2d, a contradiction. Therefore dP_2a. But dP_2a implies $dWaWbWcWd$. Since semiorders are irreflexive, dI_1d implies aI_1b, a contradiction. Therefore $\sim bWc$.

Case 2. cWb. But $\sim cP_1b$, so cI_1b, for bP_1c implies bWc, which we know is impossible. Since W is a weak order defined on A, either aWc or cWa. If cWa, then $cWaWb$ and cI_1b, so aI_1b, a contradiction. Therefore aWc. But $aWcWd$ and $\sim cI_2d$ implies $\sim aI_2d$. Now $\sim aI_2d$ and $\sim aP_2d$ implies dP_2a. Thus $aWcWdWa$. Now aI_2a implies dI_2a, a contradiction. Q.E.D.

We know that if (A, P_1, P_2) satisfies the strong double semiorder condition and (A, P_1) and (A, P_2) are both semiorders, then (A, P_1) and (A, P_2) are upper interval homogeneous. We next show that even in the presence of the assumption that $s(P_1, P_2) > m(P_1, P_2)$, the converse is false. We show the following (which by Theorem 8 is stronger than what we have just claimed): there is a nested pair of homogeneous semiorders (A, P_1, P_2) with $s(P_1, P_2) > m(P_1, P_2)$ which violates the no-balanced-cycle condition, or equivalently, is not representable. As an example consider the system (A, P_1, P_2) depicted in Fig. 2. (A, P_1, P_2) is a nested family of homogeneous semiorders. Homogeneity follows because a, d, b, c is a common compatible weak order. Now $m(P_1, P_2) = 1$ and $s(P_1, P_2) = 2$ so $m(P_1, P_2) < s(P_1, P_2)$. Yet (A, P_1, P_2) is not representable. For we would have $f(a) > f(b) + \delta_1 > f(c) + \delta_1 + \delta_2$ and $f(d) \leqq f(c) + \delta_1$ and $f(a) \leqq f(d) + \delta_2$. Therefore $f(a) \leqq f(c) + \delta_1 + \delta_2$ also, a contradiction.

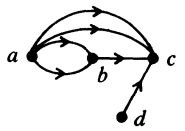


FIG. 2

9. Double indifference graphs. The results of § 8 can be translated into characterizations of double indifference graphs analogous to Corollary 1.2 of Theorem 1.

A *multigraph* (G_1, G_2) is a pair of graphs with the same vertex set. We say (G_1, G_2) is a *double indifference graph* if there exist real numbers $\delta_1 > \delta_2 > 0$ so that (5) holds for some f, and a *strong double indifference graph* if for every $\delta_1 > \delta_2 > 0$, (5) holds for some f. By definition of indifference graph, if (G_1, G_2) is a double indifference graph, then G_1 and G_2 are indifference graphs. As in Corollary 1.2, we have:

THEOREM 9. *A multigraph (G_1, G_2) is a double indifference graph if and only if there exists a balanced double semiorder system (V, P_1, P_2) with $V = V(G_1) = V(G_2)$, such that P_1 and P_2 are orientations of $\tilde{G}_1$ and $\tilde{G}_2$, respectively, and such that $m(P_1, P_2) < s(P_1, P_2)$.*

COROLLARY 9.1. *A multigraph (G_1, G_2) is a strong double indifference graph if and only if there exists a balanced double semiorder system (V, P_1, P_2) with $V = V(G_1) = V(G_2)$, such that P_1 and P_2 are orientations of $\tilde{G}_1$ and $\tilde{G}_2$, respectively, and such that $m(P_1, P_2) \leqq 1$ and $s(P_1, P_2) = \infty$.*

We illustrate these results by considering the multigraphs of Figs. 3, 4 and 5. In drawing these multigraphs, we note that since $\delta_1 > \delta_2$, representability implies that $E(G_2) \subseteq E(G_1)$. Hence, we must have a nested pair of indifference graphs, and we can represent them unambiguously as a multigraph by including two edges between x and y if $\{x, y\} \in E(G_1)$ and $E(G_2)$, and one edge if $\{x, y\} \in E(G_1)$ but not $E(G_2)$.

Example 1. (G_1, G_2) as shown in Fig. 3 is a strong double indifference graph. Choose any $\delta_1 > \delta_2 > 0$. Then the function f such that $f(a) = 0$, $f(b) = \delta_2$, and $f(c) = \delta_1 + \delta_2$ satisfies (5).

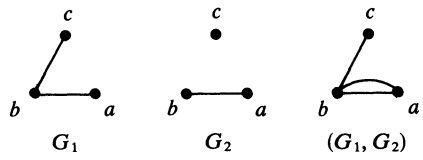


FIG. 3

Example 2. (H_1, H_2) as shown in Fig. 4 is a double indifference graph, but not a strong double indifference graph. Let $\delta = 1.5$. Then the function f such that $f(a) = 2.4$, $f(b) = 1$, $f(c) = 0$, and $f(d) = -1$ satisfies (5) with δ and 1, so (H_1, H_2) is a double indifference graph. But no such function exists with δ and 1 if $\delta > 2$, since we must have $|f(b) - f(c)| \leqq 1$ and $|f(c) - f(d)| \leqq 1$, thus $|f(b) - f(d)| \leqq 2$. Yet since no edge $\{b, d\}$ exists, $|f(b) - f(d)| > \delta$. Therefore $1 < \delta < 2$ for a function f to exist satisfying (5).

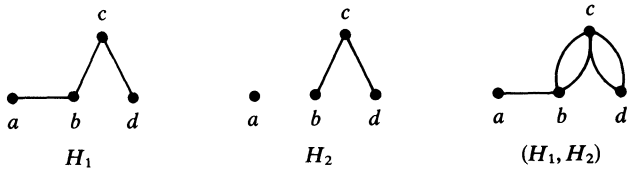


FIG. 4

Example 3. (K_1, K_2) as shown in Fig. 5 is not a double indifference graph. Suppose a function f and constant $\delta > 1$ exist satisfying (5) with δ and 1. By (5) the function f induces a weak order on A *compatible* with both K_1 and K_2. That is, we have $f(x) \leqq f(y) \leqq f(z)$ and $\{x, z\} \in E(K_i)$ implies $\{x, y\} \in E(K_i)$ and $\{y, z\} \in E(K_i)$. But the only orders compatible with K_2 have a coming first or last. In any such order, either b is between a and c or c is between a and b. But the order is also compatible with K_1. Then aI_1c or aI_1b implies bI_1c, a contradiction. Therefore no constant $\delta > 1$ and no function f exist satisfying (5) for (K_1, K_2).

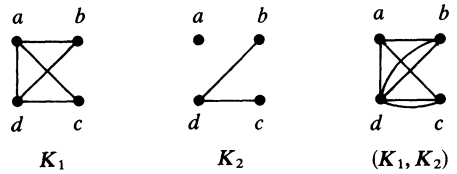


FIG. 5

To explain these results from our theorems, note that each of the graphs G_1, G_2, H_1, H_2, K_1, K_2 in Figs. 3, 4 and 5 is an indifference graph, as is easy to check. However, if we look at the complements of each of (G_1, G_2), (H_1, H_2) and (K_1, K_2), as illustrated in Fig. 6, we see some obvious differences. (Note that complement is in the multigraph sense—we consider a maximum of 2 edges possible between any two vertices.) The set $\{b, d, c\}$ in $(\widetilde{H_1, H_2})$ forms an $(I_2 - P_1)$-cycle (as defined in § 8) regardless of the orientation P_1 of $\{b, d\}$, since we have the cycle $b\bar{P}_2^* c\bar{P}_2^* dP_1 b$ or the cycle $d\bar{P}_2^* c\bar{P}_2^* bP_1 d$. Thus there exists a cycle C such that $s_C(P_1, P_2) = 2$. Therefore by Theorem 7, for any orientations P_1 and P_2 of $\tilde{H}_1$ and $\tilde{H}_2$ respectively, if there exists a function $f: V \to \mathbb{R}$ satisfying (4), δ_1/δ_2 must be less than $s_C(P_1, P_2) = 2$. Thus, we see why (H_1, H_2) is not a strong double indifference graph.

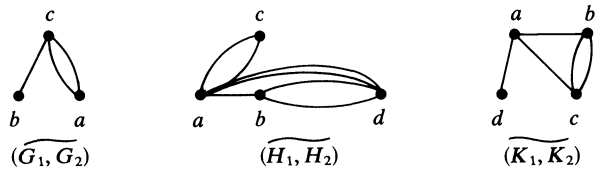


FIG. 6

Next, any orientation R_2 of $\tilde{K}_2$ must orient all edges out of a or all edges into a for R_2 to be a semiorder. Without loss of generality consider R_2 orienting all edges out of a. The two possible orientations are illustrated in Fig. 7. If R_1 is an orientation of $\tilde{K}_1$ so that $R_1 \subseteq R_2$, then $R_1 = \{(b, c)\}$ in Fig. 7(a) and $R_1 = \{(c, b)\}$ in Fig. 7(b). The pair (R_1, R_2) is shown in these two cases in Figs. 8(a) and 8(b), respectively. In Fig. 8(a), aR_2bR_1c and $\sim aR_1d$ and $\sim dR_2c$, so $aR_2bR_1c\bar{R}_2^*d\bar{R}_1^*a$ is a balanced cycle and so (V, R_1, R_2) is not a balanced double semiorder system. In Fig. 8(b), aR_2cR_1b and $\sim aR_1d$ and $\sim dR_2b$. Hence $aR_2cR_1b\bar{R}_2^*d\bar{R}_1^*a$ is a balanced cycle, so (V, R_1, R_2) is not a balanced double semiorder system. In sum, (K_1, K_2) has no balanced double semiorder system orientation.

Finally, the following orientations of $\tilde{G}_1$ and $\tilde{G}_2$ are both semiorders and the system (V, S_1, S_2) is a balanced double semiorder system: $S_1 = \{(c, a)\}$, $S_2 = \{(c, a), (c, b)\}$. See Fig. 9 for an illustration of (S_1, S_2). Any cycle can be obtained by combining the following cycles or adding $x\bar{S}_i^*x$ for $i = 1, 2$ and $x \in \{a, b, c\}$: $cS_2b\bar{S}_1^*c$, $cS_1a\bar{S}_2^*b\bar{S}_1^*c$, $cS_1a\bar{S}_1^*b\bar{S}_1^*c$, $cS_2a\bar{S}_2^*b\bar{S}_1^*c$ and $cS_2a\bar{S}_1^*b\bar{S}_1^*c$. Thus no cycle is balanced. Also, from this analysis $m(S_1, S_2) = 1$ and $s(S_1, S_2) = \infty$. Thus (V, S_1, S_2) is a balanced double semiorder system with $m(S_1, S_2) = 1$ and $s(S_1, S_2) = \infty$. Corollary 7.1 implies that (G_1, G_2) is a strong double indifference graph.

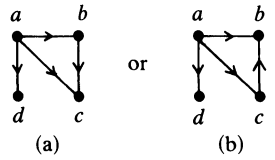


FIG. 7

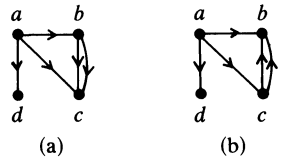


FIG. 8

FIG. 9

10. Open questions. The following questions remain open: Generalize these results to n thresholds. Find a characterization of double indifference graphs which is stated purely in terms of requirements on the graphs, e.g., in terms of forbidden subgraphs of the multigraph, rather than in terms of orienting the complements. Characterize graphs of (unit) sphericity at most 2 or at most 3 (see § 2 for definitions). Consider the generalization of double indifference graphs to pairs of graphs of (unit) sphericity at most 2 or at most 3.

REFERENCES

E. W. ADAMS AND R. F. FAGOT [1956], *A model of riskless choice*, Report 4, Applied Mathematics and Statistics Laboratory, Stanford Univ., Stanford, CA.

M. B. COZZENS [1981], *Higher and multi-dimensional analogues of interval graphs*, Ph.D. thesis, Department of Mathematics, Rutgers Univ., New Brunswick, NJ.

M. B. COZZENS AND F. S. ROBERTS, *T-colorings of graphs and the channel assignment problem*, XIIIth Southeastern Conference on Combinatorics, Graph Theory, and Computing, to appear.

A. DUCAMP [1978], *A note on an alternative proof of the representation theorem for bi-semiorder*, J. Math. Psychol., 18, pp. 100–104.

A. DUCAMP AND J. C. FALMAGNE [1969], *Composite measurement*, J. Math. Psychol., 6, pp. 359–390.

P. C. FISHBURN [1970], *Intransitive indifference with unequal indifference intervals*, J. Math. Psychol., 7, pp. 144–149.

M. GRÖTSCHEL, L. LOVÁSZ AND A. SCHRIJVER [1980], *The ellipsoid method and its consequences in combinatorial optimization*, Research Report 80151-OR, Institut für Econömetrie und Operations Research, Universität Bonn.

L. GUTTMAN [1944], *A basis for scaling qualitative data*, Amer. Sociol. Rev., 9, pp. 139–150.

W. K. HALE [1980], *Frequency assignment: theory and applications*, Proc. IEEE, 68, pp. 1497–1514

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

T. HAVEL [1982], *The combinatorial distance geometry approach to the calculation of molecular conformation*, Ph.D. Thesis, Group in Biophysics. Univ. of California, Berkeley, CA.

W.-L. HSU [1980], *Efficient algorithms for some packing and covering problems on graphs*, Ph.D. Thesis, School of Operations Research and Industrial Engineering, Cornell Univ., Ithaca, NY.

L. HUBERT [1974], *Some applications of graph theory and related non-metric techniques to problems of approximate seriation: The case of symmetric proximity measures*, British J. Math. Statist. Psychol., 27, pp. 133–153.

D. G. KENDALL [1963], *A statistical approach to Flinders Petrie's sequence dating*, Bull. Inst. Internat. Statist., 40, pp. 657–680.

———, [1969a], *Incidence matrices, interval graphs and seriation in archaeology*, Pacific J. Math., 28, pp. 565–570.

———, [1969b], *Some problems and methods in statistical archaeology*, World Archaeology, 1, pp. 61–76.

———, [1971a], *Abundance matrices and seriation in archaeology*, Z. Wahrsch. Verw. Gebiete, 17, pp. 104–112.

———, [1971b], *A mathematical approach to seriation*, Philos. Trans. Roy. Soc. London Ser. A, 269, pp. 125–135.

———, [1971c], *Seriation from abundance matrices*, in Mathematics in the Archaeological and Historical Sciences, F. R. Hodson et al., eds., Edinburgh Univ. Press, Edinburgh.

C. H. KRAFT, J. W. PRATT AND A. SEIDENBERG [1959], *Intuitive probability on finite sets*, Ann. Math. Statist., 30, pp. 408–419.

D. H. KRANTZ, R. D. LUCE, P. SUPPES AND A. TVERSKY, *Foundations of Measurement*, Vol. II, Academic Press, New York, to appear.

R. D. LUCE [1956], *Semiorders and a theory of utility discrimination*, Econometrica, 24, pp. 178–191.

I. RABINOVITCH [1978], *The dimension of semiorders*, J. Combin. Theory, Ser. A, 25, pp. 50–61.

F. S. ROBERTS [1968], *Representations of indifference relations*, Ph.D. Thesis, Department of Mathematics, Stanford Univ., Stanford, CA.

———, [1969], *Indifference graphs*, in Proof Techniques in Graph Theory, F. Harary, ed., Academic Press, New York, pp. 139–146.

———, [1971a], *Homogeneous families of semiorders and the theory of probabilistic consistency*, J. Math. Psychol., 8, pp. 248–263.

F. S. ROBERTS [1971b], *On the compatibility between a graph and a simple order*, J. Combin. Theory, 11, pp. 28–38.

———, [1976], *Discrete Mathematical Models, with Applications to Social, Biological, and Environmental Problems*, Prentice-Hall, Englewood Cliffs, NJ.

———, [1978], *Graph Theory and its Applications to Problems of Society*, CBMS Regional Conference Series in Applied Mathematics 29, Society for Industrial and Applied Mathematics, Philadelphia.

———, [1979a], *Indifference and seriation*, in Advances in Graph Theory, F. Harary, ed., Proc. New York Acad. Sci., 328, pp. 171–180.

———, [1979b], *Measurement Theory, with Applications to Decisionmaking, Utility, and the Social Sciences*, Addison-Wesley, Reading, MA.

D. SCOTT [1964], *Measurement models and linear inequalities*, J. Math. Psychol., 1, pp. 233–247.

D. SCOTT AND P. SUPPES [1958], *Foundational aspects of theories of measurement*, J. Symbolic Logic, 23, pp. 113–128.

J. A. ZOELLNER AND C. L. BEALL [1977], *A breakthrough in spectrum conserving frequency assignment technology*, IEEE Trans. Electromag. Comput., EMC-19, pp. 313–319.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0019 $01.00/0

ON THE GREEDY HEURISTIC FOR CONTINUOUS COVERING AND PACKING PROBLEMS*

MARSHALL L. FISHER† AND LAURENCE A. WOLSEY‡

Abstract. Worst-case bounds are given on the performance of the greedy heuristic for a continuous version of the set covering problem. This generalizes results of Chvatal, Johnson and Lovasz for the 0–1 covering problem. The results for the greedy heuristic and for other heuristics are obtained by treating the covering problem as a limiting case of a generalized location problem for which worst-case results are known. An alternative approach involving dual greedy heuristics leads also to worst-case bounds for continuous packing problems.

Introduction. This paper deals with a worst-case study of the greedy heuristic for the continuous covering problem

$$(C(b)) \qquad \begin{aligned} Z(b) &= \min cy, \\ Ay &\geqq b, \\ y &\geqq 0, \end{aligned}$$

where A is an $m \times n$ matrix, and A, b and c are nonnegative rationals. Although $(C(b))$ contains no integrality restriction on y, the greedy heuristic always produces an integer/0–1 solution when A is a 0–1 matrix and b is integer/$(1, \cdots, 1)^T$. Therefore our results can be seen as a generalization of those of Johnson [5], Lovasz [6] and Chvatal [1] who considered the performance of the greedy heuristic when A is 0–1, $b = (1, \cdots, 1)^T$, and y is required to be 0 or 1.

A novel feature of our analysis is the use of a worst-case bound for a generalized location problem to derive the bound for the covering problem. Even though direct proofs exist, we feel that this is of interest because of the definite lack of a unified theory of heuristics; see [4], [8].

This is also one of the few analyses of a linear programming (LP) heuristic of which we are aware. LP heuristics might be useful in LP crashing procedures, and for fathoming in branch and bound. The results below also give an indication of appropriate row scaling factors for problem $C(b)$.

Section 1 states our results on greedy covering, and § 2 contains the proofs of these results. In § 3 alternative statements of our earlier results lead also to an analysis of heuristic solutions to the dual of $C(b)$, or in other words to a heuristic analysis for continuous packing problems. Finally the application of other location heuristics to covering problems is discussed.

1. Results on greedy. In the description of the greedy heuristic below we use the notation

$$s_i(y) = \max\left(0, b_i - \sum_{j=1}^{n} a_{ij}y_j\right) \quad \text{and} \quad \lambda_j^{t-1} = \left(\sum_{i \in M^{t-1}} a_{ij}\right) \Big/ c_j.$$

A greedy heuristic for $C(b)$

(1) *Initialization*: Set $y^0 = 0, M^0 = \{1, \cdots, m\}, s_i(y^0) = b_i, i = 1, \cdots, m$, and calculate $\lambda_j^0, j = 1, \cdots, n$. Set $t = 1$.

* Received by the editors June 4, 1981, and in revised form June 3, 1982.

† Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104.

‡ Center for Operations Research and Econometrics, Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium.

(2) *Step t*:
   a) Find $\theta^t = \min_j (1/\lambda_j^{t-1})$, and choose a variable j_t for which $\theta^t = c_{j_t}/\sum_{i \in M^{t-1}} a_{ij_t}$.
   b) Find $\eta^t = \min_{i \in M^{t-1}} (s_i(y^{t-1})/a_{ij_t})$, and choose a row $i_t \in M^{t-1}$ for which $\eta^t = s_{i_t}(y^{t-1})/a_{i_t j_t}$.
   c) Set $M^t = M^{t-1} - \{i_t\}$, and $y^t = y^t + \eta^t e_{j_t}$, where e_{j_t} is the j_tth unit vector.

(3) If $t < m$, set $t = t+1$. Update $s_i(y^t)$, $i \in M^t$ and λ_j^t, $j = 1, \cdots, n$, and go to 2. If $t = m$, set $y^G = y^m$, $Z^G(b) = cy^G$ and stop.

If there are ties in the selection of i_t, one could reduce execution time of greedy by removing more than one element from M^{t-1}. However, forcing greedy to execute step 2 exactly m times as we have done above will simplify the proofs. In addition it allows us to assume without loss of generality that $i_t = t$, $t = 1, \cdots, m$, and hence $M^{t-1} = \{t, \cdots, m\}$. Note that the running time of the heuristic is $O(m \cdot \max(m, n))$ as it requires at most $2(m+n)$ multiplications and divisions at each of the m steps.

For later use we also introduce the following m vectors $\{u^t\}_{t=1}^m$ defined by $u_i^t \theta^t$ for $i = t, \cdots, m$ and $u_i^t = 0$ otherwise, and $u^* = (\theta^1, \cdots, \theta^m)$.

Note that the choices made by greedy can be affected by scaling the rows of $C(b)$, but are unaffected by column scaling. Row scaling is of major importance in the results given below.

To specify the first result, let

$$\alpha = \min_i \max_j \left(\frac{a_{ij}}{c_j}\right) \leqq \frac{1}{\theta^m} \quad \text{and} \quad \beta = \max_j \left(\sum_{i=1}^m \frac{a_{ij}}{c_j}\right) = \frac{1}{\theta^1}.$$

THEOREM 1. *If $Z^G(b)$ is the value of a greedy heuristic solution for $C(b)$, then*

$$Z^G(b) \leqq \left[1 + \log_e\left(\frac{\beta}{\alpha}\right)\right] Z(b).$$

Minimizing the value of β/α leads naturally to a scaling rule:

Canonical form 1. $\max_j (a_{ij}/c_j) = \alpha$ for all $i = 1, \cdots, m$.

Note that if d is the maximum number of nonzero elements in any column of A, and canonical form 1 is adopted, then $\beta/\alpha \leqq d$, so we obtain the following corollary.

COROLLARY. *If $C(b)$ is in canonical form* 1, *then*

$$Z^G(b) \leqq (1 + \log_e d) Z(b).$$

The appearance of term β/α in the result of Theorem 1 is not surprising if we note that it is an upper bound on the ratio of θ^m/θ^1, i.e., the ratio of the most expensive to the cheapest unit cost of covering units of b during the application of the greedy heuristic.

An alternative and natural scaling is obtained by simply normalizing the requirements vector b:

Canonical form 2. $b_i = 1$ for all $i = 1, \cdots, m$.

THEOREM 2. *If $C(b)$ is in canonical form* 2, *then*

$$Z^G(b) \leqq (1 + \log_e m) Z(b).$$

Now consider the integer covering problem with A 0–1 and b integer. In this case it is clear that the greedy heuristic for $C(b)$ will produce an integer solution, and we obtain an integer programming heuristic with a bound for $Z^G(b)/Z(b)$ of $(1 + \log_e d)$ with canonical form 1, and of $(1 + \log_e m)$ with canonical form 2. When $b = (1, \cdots, 1)$, Chvatal has a tight bound of $\sum_{i=1}^d (1/i)$ with canonical form 2, and recently Dobson

[3] has obtained an important generalization of this result, so it is still an open question whether Theorem 2 can be strengthened to $(1+\log_e d)$ for general $C(b)$. The worst-case examples of Lovasz [6] show that Theorems 1 and 2 are tight asymptotically.

2. Proof. This section contains the proofs of Theorems 1 and 2. Both make use of two observations: i) a close connection exists between $C(b)$ and a generalized location problem $L(\lambda, b)$ described below, and ii) the behavior of the greedy heuristic for $L(\lambda, b)$ can be described in invoking a result from [9].

Without loss of generality we shall assume that $C(b)$ has been column-scaled so that $c_j = 1$ for all $j = 1, \cdots, n$.

Consider the problem:

$$
\begin{aligned}
W(\lambda, b) = \max \sum_{j=1}^{n} \sum_{i=1}^{m} a_{ij} x_{ij}, & \\
\sum_{j=1}^{n} a_{ij} x_{ij} \leqq b_i, \qquad & i = 1, \cdots, m, \\
0 \leqq x_{ij} \leqq y_j \leqq h_j, \qquad & i = 1, \cdots, m, \quad j = 1, \cdots, n, \\
\sum_{j=1}^{n} y_j \leqq \lambda & \qquad (L(\lambda, b))
\end{aligned}
$$

with a_{ij}, b_i and h_j nonnegative rationals for all i and j.

The greedy heuristic for $L(\lambda, b)$, given below, with $h_j = +\infty$ for all j, is exactly the same as the greedy heuristic for $C(b)$ with $c_j = 1$ for all j except that it may terminate earlier. In other words, the y^t vectors generated at each step are identical.

A greedy heuristic for $L(\lambda, b)$

(1) *Initialization*: Set $y^0 = 0, M^0 = \{1, \cdots, m\}, s_i(y^0) = b_i, i = 1, \cdots, m$, and calculate $\lambda_j^0, j = 1, \cdots, m$. Set $t = 1$.

(2) *Step* t:

a) Find $\theta^t = \min_j (1/\lambda_j^{t-1})$, and choose a variable j_t for which $\theta^t = c_{j_t}/\sum_{i \in M^{t-1}} a_{ij_t}$.

b) Find $\eta^t = \min_{i \in M^{t-1}} s_i(y^{t-1})/a_{ij_t}$, and choose a row $i_t \in M^{t-1}$ for which $\eta^t = s_{i_t}(y^{t-1})/a_{i_t j_t}$. If $\sum_{j=1}^n y_j^{t-1} + \eta^t > \lambda$, set $y^t = y^{t-1} + (\lambda - \sum_{j=1}^n y_j^{t-1})\, e_{j_t}$, and go to step (2d).

c) Set $M^t = M^{t-1} - \{i_t\}$, and $y^t = y^{t-1} + \eta^t e_{j_t}$.

d) Set $x_{ij_t}^t = \min \{y_{j_t}, [b_i - \sum_{j \neq j_t} a_{ij} x_{ij}^{t-1}]/a_{ij_t}\}$ and $x_{ij}^t = x_{ij}^{t-1}$ for $j \neq j_t$, for all $i = 1, \cdots, m$.

(3) If $\sum_{j=1}^n y_j^t < \lambda$, set $t = t+1$, update $s_i(y^t)$, $i \in M^t$ and λ_j^t, $j = 1, \cdots, n$ and go to (2). Otherwise set $(x^G, y^G) = (x^t, y^t)$, $W^G(\lambda, b) = \sum_{j=1}^n \sum_{i=1}^m a_{ij} x_{ij}^G$, and stop.

It is easily verified that

$$
w(y) = \max \left\{ \sum_{j=1}^{n} \sum_{i=1}^{m} a_{ij} x_{ij} : \sum_{j=1}^{n} a_{ij} x_{ij} \leqq b_i,\ i = 1, \cdots, m, \right.
$$
$$
\left. 0 \leqq x_{ij} \leqq y_j,\ i = 1, \cdots, m, j = 1, \cdots, n \right\}
$$

is submodular on $\mathbb{R}_+^n$. This enables us to apply, [9, Thm. 1] and establish the following result:

THEOREM 3. *Let* $W^G(\lambda, b)$ *denote the objective value of a greedy solution for* $L(\lambda, b)$. *Then*

$$
W^G(\lambda, b) \geqq (1 - e^{-\lambda/\mu}) W(\mu, b) \quad \forall \lambda, \mu \geqq 0.
$$

If all a_{ij} are 0 or 1, all $b_i = 1$, and $\lambda = \mu$ is integer, then $L(\lambda, b)$ is the λ-median problem with 0–1 costs. A greedy solution will satisfy $y_j = 0$ or 1 for all j and Theorem 3 is equivalent to a result given by Cornuéjols, Fisher and and Nemhauser [2].

Proof of Theorem 1. We shall need three conditions that follow from the close connection between problems $C(b)$ and $L(\lambda, b)$.

Condition A. $W(\lambda, b) = \sum_{i=1}^{m} b_i$ if and only if $\lambda \geqq Z(b)$.

Suppose $W(\lambda, b) = \sum_{i=1}^{m} b_i$, and let (x, y) be optimal for $L(\lambda, b)$. Then it follows that $\sum_{j=1}^{n} a_{ij} x_{ij} = b_i$, $i = 1, \cdots, m$ and $\sum_{j=1}^{n} y_j \leqq \lambda$. As $0 \leqq x_{ij} \leqq y_j$, we conclude that $\sum_{j=1} a_{ij} y_j \geqq b_i$, $i = 1, \cdots, m$, and hence y is feasible in $C(b)$, and $\sum_{j=1}^{n} y_j \geqq Z(b)$. Conversely, if y is optimal in $C(b)$, and $\lambda \geqq Z(b) = \sum_{j=1}^{n} y_j$, it is easy to construct x such that (x, y) is optimal for $L(\lambda, b)$ with $W(\lambda, b) = \sum_{i=1}^{m} b_i$.

The next two conditions depend on the fact that if the greedy heuristic is applied to $L(\lambda, b)$ or to $C(b)$ the greedy solution produced is identical.

Condition B. $W^G(\lambda, b) = \sum_{i=1}^{m} b_i$ if and only if $\lambda \geqq Z^G(b)$, provided an identical tie-breaking rule is used for each heuristic.

Condition C. Let b^R be the part of b still not covered after applying the greedy heuristic to $L(\lambda, b)$. Then $Z^G(b) = \lambda + Z^G(b^R)$.

Now let $\lambda^* = Z(b) \log_e (\beta/\alpha)$, and consider problem $L(\lambda^*, b)$. If $W^G(\lambda^*, b) = \sum_{i=1}^{m} b_i$, we are done, as $1 + Z(b) \log_e (\beta/\alpha) > \lambda^* \geqq Z^G(b)$, where the last inequality follows from Condition B.

If not, let b^R be the remainder as defined above when $\lambda = \lambda^*$. By Condition C, $Z^G(b) = Z(b) \log_e (\beta/\alpha) + Z^G(b^R)$, and it only remains to show that $Z^G(b^R) \leqq Z(b)$.

As $\alpha = \min_i \min_j a_{ij}$ was defined so that $1/\alpha$ is the worst price that the greedy heuristic pays per unit of b, $\theta^t \leqq 1/\alpha$ for all iterations t of the greedy heuristic, and hence $Z^G(b^R) \leqq (1/\alpha) \sum_{i=1}^{m} b_i^R$. Now

$$\sum_{i=1}^{m} b_i - \sum_{i=1}^{m} b_i^R = W^G(\lambda^*, b) \geqq (1 - e^{-\log_e (\beta/\alpha)}) W(Z(b), b) = \left(1 - \frac{\alpha}{\beta}\right) \sum_{i=1}^{m} b_i,$$

where the first equality follows from the definition of b^R, the inequality from Theorem 3, and the last equality from Condition A. Hence

$$\frac{1}{\alpha} \sum_{i=1}^{m} b_i^R \leqq \frac{1}{\beta} \sum_{i=1}^{m} b_i.$$

Finally, since $c_j = 1$ for all j, and $\beta = \max_j \sum_{i=1}^{m} a_{ij}$, $(1/\beta, \cdots, 1/\beta)$ is feasible in the dual of $C(b)$, and hence $(1/\beta) \sum_{i=1}^{m} b_i \leqq Z(b)$, and $Z^G(b^R) \leqq Z(b)$ as required. □

Proof of Theorem 2. Using an identical argument with $b = (1, \cdots, 1)^T$, we let $\lambda^* = Z(b) \log_e m$, and let b^R be the remainder after applying greedy to $L(\lambda^*, b)$, so that $Z^G(b) = Z(b) \log_e m + Z^G(b^R)$.

Now by Theorem 3, $W^G(\lambda^*, b) \geqq (1 - e^{-\log_e m}) W(Z(b), b) = (1 - 1/m) \sum_{i=1}^{m} b_i = m - 1$, and hence $\sum_{i=1}^{m} b_i^R \leqq 1$. Then using $b_m = 1$, $Z^G(b^R) \leqq \theta^m \sum_{i=1}^{m} b_i^R \leqq \theta^m b_m \leqq Z(b)$ as $\theta^m b_m$ is the optimal value for problem $C(b)$ with its first $(m-1)$ constraints removed. □

3. Covering and packing heuristics. Consider again the covering problem $C(b)$, and its dual, the packing problem:

$$(P(b)) \qquad Z(b) = \max \{ub : uA \leqq c, u \geqq 0\},$$

and the vectors $\{u^t\}_{t=1}^{m}$ and u^* defined during the greedy heuristic. Let $W_1^H = u^* b / \{1 + \log_e (\beta/\alpha)\}$ be the value of *dual greedy heuristic* 1 and $W_2^H = \cdot max_{t=1,\cdots,m} u^t b$ be the value of *dual greedy heuristic* 2.

THEOREM 4. *If the greedy algorithm is applied to $C(b)$, then*
a) *the vector $(\theta^1, \cdots, \theta^m)/\{1+\log_e(\beta/\alpha)\}$ is feasible in $P(b)$, and*

$$W_1^H \geqq \left\{1+\log_e\left(\frac{\beta}{\alpha}\right)\right\}^{-1} Z^G(b);$$

b) *each of the vectors $u^t = (0, \cdots, 0, \theta^t, \cdots, \theta^t)$ is feasible in $P(b)$.*
If $C(b)$ is in canonical form 2,

$$W_2^H \geqq \left(\sum_{i=1}^{m} \frac{1}{i}\right)^{-1} Z^G(b).$$

Proof. a) Let $k = (1+\log_e(\beta/\alpha))$. We claim first that $(\theta^1, \cdots, \theta^m)/k$ is dual feasible for $C(b)$. For each j, we have from the greedy algorithm that

$$\begin{aligned}
&\theta^1(a_{1j}+a_{2j}+\cdots+a_{mj}) \leqq 1,\\
&\theta^2(\quad a_{2j}+\cdots+a_{mj}) \leqq 1,\\
&\qquad\vdots\\
&\theta^m(\qquad\qquad a_{mj}) \leqq 1
\end{aligned}$$

with $0<\theta^1 \leqq \theta^2 \leqq \cdots \leqq \theta^m$, and $\theta^m/\theta^1 \leqq \beta/\alpha$. The claim is proven if we can show that $\zeta_j = \theta^1 a_{1j} + \cdots + \theta^m a_{mj} \leqq k$ for all j.

Note that for all values of a_{ij} and θ^t satisfying these constraints,

$$\begin{aligned}
\zeta_j &= (a_{1j}+\cdots+a_{mj})\theta^1 + (a_{2j}+\cdots+a_{mj})(\theta^2-\theta^1)+\cdots+a_{mj}(\theta^m-\theta^{m-1})\\
&\leqq \theta^1 + \frac{\theta^2-\theta^1}{\theta^2} + \frac{\theta^3-\theta^2}{\theta^3} + \cdots + \frac{\theta^m-\theta^{m-1}}{\theta^m}\\
&= m - \frac{\theta^1}{\theta^2} - \frac{\theta^2}{\theta^3} - \cdots - \frac{\theta^{m-1}}{\theta^m}.
\end{aligned}$$

Letting $\rho_t = \theta^t/\theta^{t+1}$, we obtain an upper bound on ζ_j by calculating

$$\min\left\{\sum_{t=1}^{m-1} \rho_t : \prod_{t=1}^{m-1} \rho_t \geqq \frac{\alpha}{\beta}, \rho_t > 0, t = 1, \cdots, m-1\right\},$$

whose value is well known to be $(m-1)^{m-1}\sqrt{\alpha}/\beta$. Hence

$$\begin{aligned}
\zeta_j &\leqq 1+(m-1)\{1-(\beta/\alpha)^{-1/(m-1)}\} \leqq 1+(m-1)\left\{1-1+\frac{1}{m-1}\log_e(\beta/\alpha)\right\}\\
&= 1+\log_e(\beta/\alpha) \text{ as } e^{-x} \geqq 1-x \quad \forall x \geqq 0.
\end{aligned}$$

Now

$$\begin{aligned}
kW_1^H = u^*b = \sum_{i=1}^{m} \theta^i b_i &= \sum_{i=1}^{m} \theta^i \sum_{t=1}^{i} a_{ij_t} y_{j_t} \geqq \sum_{i=1}^{m}\sum_{t=1}^{i} \theta^t a_{ij_t} y_{j_t}\\
&= \sum_{t=1}^{m} y_{j_t} \sum_{i=t}^{m} \theta^t a_{ij_t} = \sum_{t=1}^{m} c_{j_t} y_{j_t} = Z^G(b),
\end{aligned}$$

and a) is proved.

b) From the greedy algorithm and the definition of θ^t, $\theta^t(\sum_{i=t}^{m} a_{ij}) \leqq 1$ for all j. In other words, u^t is feasible in $P(b)$ for $t = 1, \cdots, m$, with dual value $u^t b = (m-t+1)\theta^t$, when $b_i = 1$ for all i.

Above it is shown that $Z^G(b) \leqq \sum_{t=1}^{m} \theta^t b_t$. Now as $b_i = 1$ for all i and

$$W_2^H = \max_t (m-t+1)\theta^t, \qquad Z^G(b) \leqq \sum_{t=1}^{m} \frac{W_2^H}{m-t+1} \leqq W_2^H \sum_{t=1}^{m} \frac{1}{i}. \qquad \square$$

Note that if we start from a continuous packing problem in standard form,

$$(PP) \qquad W = \max\{cx : Ax \leqq b, x \geqq 0\},$$

the dual greedy heuristic can be obtained directly by applying the "equality heuristic" described below.

An equality heuristic for (PP)
Set $t = 1$.
Iteration t:

a) Let $\alpha^t = \max\{\alpha : \alpha(\sum_{j=t}^{n} a_{ij}) \leqq b_i \text{ for all } i = 1, \cdots, m\}$ and suppose the maximum is determined by some row i_t, i.e.,

$$\alpha^t\left(\sum_{j=t}^{n} a_{i_t j}\right) = b_{i_t}.$$

b) Find which variable $j_t \in \{t, \cdots, m\}$ is least profitable when considering only constraint i_t, i.e.,

$$j_t = \arg\min_{j=t,\cdots,n} \frac{c_j}{a_{i_t j}}.$$

Reorder the columns so that $j_t = t$.

c) If $t = n$, stop. Otherwise set $t \leftarrow t+1$.

$x^* = (\alpha^1, \cdots, \alpha^n)/(1+\log_e(\beta/\alpha))$ is *heuristic solution* 1 with value $W_1^H = cx^*$. Define $x^t \in R^n$ by $x_j^t = 0$ for $j < t$ and $x_j^t = \alpha^t$ for $j \geqq t$. Supppose $\max_t cx^t = cx^{\tilde{t}}$. $\tilde{x} = x^{\tilde{t}}$ is *heuristic solution* 2 with value $W_2^H = c\tilde{x}$.

It should now be evident how Theorem 4 can be restated for the equality heuristic applied to the packing problem (*PP*).

4. Extensions. If we now consider the covering problem with upper bounds:

$$C(b,h) \qquad \begin{aligned} Z(b,h) = \min\ & cy, \\ & Ay \geqq b, \\ & 0 \leqq y \leqq h, \end{aligned}$$

and adapt the greedy heuristic to incorporate the upper bounds on y, an identical analysis to that of Theorem 1 using Theorem 3 gives:

THEOREM 5. *Assume problem $C(b,h)$ is feasible, and let $\tilde{\alpha} = \min_{i,j}\{a_{ij}/c_j : a_{ij} \neq 0\}$. If $Z^G(b,h)$ is the value of a greedy heuristic solution, then*

$$Z^G(b,h) \leqq \left(1 + \log_e \frac{\beta}{\tilde{\alpha}}\right) Z(b,h).$$

Another natural extension is to consider other heuristics for the location problem, and apply them to the covering problem. Here we suppose A is 0–1, $b = (1, \cdots, 1)^T$ and $c = (1, \cdots, 1)$. Let H be any heuristic for $L(\lambda, b)$ that satisfies the following:

$$W^H(\lambda, b) \geqq (1 - e^{-\gamma}) W(\lambda, b) \quad \text{for some } \gamma > 0 \text{ and } \lambda \text{ integer}.$$

For example, the k-enumeration plus greedy heuristic given in [7] enjoys these

properties with $e^{-\gamma} = ((m-k)/m)\,e^{-1}$, and the interchange heuristic with $\gamma = \log_e 2$. Below we give a heuristic for this special case of $C(b)$ that uses H as a subroutine.

Heuristic for $C(b)$ with $A = (0, 1)$-matrix, $c = (1, \cdots, 1)$, $b = (1, \cdots, 1)^T$

Initialization: Set $k = 1$ and $s^0 = b$.

Step k: Using bisection, find an integer λ_k for which $W^H(\lambda_k, s^{k-1}) \geqq (1-e^{-\gamma})\sum_{i=1}^m s_i^{k-1}$, and $W^H(\lambda_k - 1, s^{k-1}) < (1-e^{-\gamma})\sum_{i=1}^m s_i^{k-1}$. Let (x_{ij}^k, y_j^k) denote the resulting solution and set $s_i^k = s_i^{k-1} - \sum_{j=1}^n a_{ij}x_{ij}^k$. If $s_i^k = 0$ for all i, set $y_j^G = y_j^t$ for all j, $Z^H(b) = \sum_{j=1}^n c_j y_j^G$. Otherwise set $k = k+1$ and repeat step k.

THEOREM 6.

$$Z^H(b) \leqq \left(1 + \frac{1}{\gamma}\log_e m\right) Z(b).$$

Proof. The result clearly holds for $m = 1$. Assume it holds for all problems with less than m rows.

Claim 1. $\lambda_1 \leqq Z(b)$.

As $Z(b)$ is optimal for $C(b)$, $W(\lambda, b) = m$ (for all $\lambda \geqq Z(b)$) and hence from the assumptions $W^H(Z(b), b) \geqq (1-e^{-\gamma})m$ for all applications of heuristic H. As $W^H(\lambda_1 - 1, b) < (1-e^{-\gamma})m$, $\lambda_1 - 1 < Z(b)$, and hence $\lambda_1 \leqq Z(b)$.

Claim 2. $\sum_{i=1}^m s_i^1 \leqq e^{-\gamma}m$ by definition of λ_i.

Claim 3. $Z(s^1) \leqq Z(b)$ as the optimal solution for $C(b)$ is clearly feasible in $C(s^1)$.

Now from the heuristic

$$\begin{aligned} Z^H(b) &= \lambda_1 + Z^H(s^1) \\ &\leqq Z(b) + \left(1 + \frac{1}{\gamma}\log_e\left(\sum_{i=1}^m s_i^1\right)\right) Z(s^1) \qquad \text{(by Claim 1 and induction)} \\ &\leqq Z(b) + \left(1 + \frac{1}{\gamma}\log_e (e^{-\gamma}m) Z(b)\right) \qquad \text{(by Claims 2 and 3)} \\ &= \left[1 + \frac{1}{\gamma}\log_e m\right] Z(b). \end{aligned}$$

□

Note also that this heuristic must terminate in at most $\lceil (1/\gamma)\log_e m \rceil$ steps.

REFERENCES

[1] V. CHVÁTAL, *A greedy heuristic for the set-covering problem*, Math. Oper. Res., 4 (1979), pp. 233–235.

[2] G. CORNUÉJOLS, M. L. FISHER AND G. L. NEMHAUSER, *Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms*, Management Sci, 23 (1978), pp. 789–810.

[3] G. DOBSON, *Worst case analysis of greedy heuristics for integer programming with nonnegative data*, Tech. Rep. SOL 80–25, Stanford Univ., Stanford, CA, October 1980.

[4] M. L. FISHER, *Worst-case analysis of integer programming heuristics*, Management Sci., 26 (1980), pp. 1–17.

[5] D. S. JOHNSON, *Approximation algorithms for combinatorial problems*, J. Comput. System Sci., 9 (1974), pp. 256–298.

[6] L. LOVASZ, *On the ratio of optimal integral and fractional covers*, Discrete Math., 13 (1975), pp. 383–390.

[7] G. L. NEMHAUSER, L. A. WOLSEY AND M. L. FISHER, *An analysis of approximations for maximizing submodular set functions*—I, Math. Programming, 14 (1978), pp. 265–294.

[8] L. A. WOLSEY, *Heuristic analysis, linear programming and branch and bound*, Math. Programming Study, 13 (1980), pp. 121–134.

[9] ———, *Maximixing real-valued submodular functions: Primal and dual heuristics for location problems*, CORE Discussion Paper 8019, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, 1980, revised February 1981.

[10] ———, *An analysis of the greedy algorithm for the submodular (set) covering problem*, CORE Discussion Paper 8125, Université Catholique de Louvain, August 1981.

SIAM J. ALG. DISC. METH.
Vol. 3, No. 4, December 1982

0196-5212/82/0304-0020 $01.00/0

RECURSIVE ALGORITHMS FOR UNITARY AND SYMPLECTIC GROUP REPRESENTATIONS*

KENNETH BACLAWSKI†

Abstract. The finite-dimensional irreducible representations of the unitary groups $U(n)$, $SU(n)$ and the symplectic groups $Sp(2n, \mathbb{C})$ are explicitly constructed using recursive algorithms. A simple labelling system is described that provides a unique label for each vector in a specific basis of every irreducible representation, and the algorithms act in a "combinatorial" manner on those labels. The algorithms are examples of lexicographic straightening algorithms.

Introduction. This paper is concerned with the problem of explicitly constructing the finite-dimensional irreducible representations (irreps) of Lie groups—in the strong sense of actually writing out practical algorithms for computing the action of an element of the Lie group on a vector in one of the representation spaces. The algorithms we use are called lexicographic straightening algorithms, and the particular examples we analyze here were chosen to illustrate some useful features of such algorithms. We presented both $U(n)$ and $SU(n)$, even though they have almost the same representation theory, to show how one can deal with reductive (but nonsemisimple) Lie groups. The symplectic groups illustrate a more subtle process. The class of representations obtained by restricting the action of a Lie group to that of a subgroup is called a "branching rule." Since all the irreps of $Sp(2n, \mathbb{C})$ may be obtained by "branching" from irreps of $Sl(2n, \mathbb{C})$, it is convenient to construct the algorithms for symplectic group representations by modifying the algorithms for $Sl(2n, \mathbb{C})$. The resulting algorithm is part of the full algorithm that expresses the branching of representations during restriction from $Sl(2n, \mathbb{C})$ to $Sp(2n, \mathbb{C})$.

It is useful to give some background in order to place this paper within the context of the vast literature dealing with Lie groups and their representations. There are three main ingredients in each algorithm: a "combinatorial" labelling system for enumerating a basis of weight vectors of the irreps; an algebraic structure, certain elements of which are associated with the labels, thereby giving a concrete meaning to the abstract labels; and (common to all the algorithms) the concept of a lexicographic straightening algorithm. In the following discussion all attributions are to the earliest reference known to this author.

Labels for the irreps of classical groups have been known for some time. A natural way to derive them is via the branching rules for groups in each infinite sequence. For the unitary groups these are called the Weyl branching rules (Weyl (1934)). For the orthogonal groups, they are due to Gelfand–Zetlin (1950). Finally, the branching rules for the symplectic groups are due to Hegerfeldt (1967). The use of tableaux as a labelling method was developed by Young (1927), and independently by Garnir (1950), for the symmetric group and by Weyl (1934), Hodge (1942), (1943) and Hamermesh (1962) for the unitary groups. That tableaux and branching patterns are combinatorially equivalent was apparently first noticed by Baird–Biedenharn (1963). Symplectic tableaux were first developed by King (1975) although in retrospect the Baird–Biedenharn result yields these from Hegerfeldt's branching rules. More recent work has yielded tableaux unrelated to branching rules. Symplectic tableaux have

* Received by the editors November 25, 1981 and in revised form March 16, 1982. This research was supported by the National Science Foundation under grant MCS 79-03029.

† Department of Mathematics, Haverford College, Haverford, Pennsylvania 19041.

been found by DeConcini (1979). Tableaux for the orthogonal and special orthogonal groups were found by Lancaster–Towber (1979). Finally, tableaux were constructed for all classical groups by Lakshmibai–Musili–Seshadri (1979). Apparently the only exceptional group for which tableaux are known is G_2, as discussed in Baclawski–Towber (1982).

The history of algebraic structures for which tableaux serve as labels for standard forms is much longer and more complex than the history of the labels themselves. The work that eventually culminated in the shape algebra, the algebraic structure we use to "concretize" the abstract labels, goes back at least to Schweins (1825) and includes Sylvester (1851), Young (1902, 1927), Turnbull (1929), Hodge (1942, 1943), Garnir (1950) and Doubilet–Rota–Stein (1974). Each of these found a different algebraic context within which the "shuffle relations" appear in a natural way. The first functorial, characteristic-free context, the shape algebra, was developed by Towber (1977) based on work of Higman, Beetham and Carter–Lusztig, as well as all those mentioned above. All of this so far deals with the shape algebra for $Sl(n, \mathbb{C})$ (or equivalently for $SU(n)$). For the symplectic group the shape algebra was developed by DeConcini (1979). For arbitrary semisimple Lie groups, the shape algebra was defined by Towber and developed in a series of papers: Towber (1977), (1979) and Lancaster–Towber (1979, 1982). The relations for the shape algebra were first computed in general by Kostant, whose result is cited in Lancaster–Towber (1979). For additional references see Lancaster–Towber (1979).

The concept of a lexicographic straightening algorithm emerged slowly from a huge variety of special cases, especially algebraic structures discussed above, although in retrospect the essential idea already appears in Macaulay (1927). For more recent discussions see Baclawski (1981) and DeConcini–Eisenbud–Procesi (1981). For the unitary groups, the straightening algorithm we discuss is due to the list of authors enumerated in the last paragraph above. A different algorithm, which yields closed-form formulas for both unitary and orthogonal groups, is due to Gelfand–Zetlin (1950). These formulas give the action of certain generators of these groups on a basis that differs from the one we use but which uses equivalent labels. In addition the representation matrices they obtain are unitary. Our algorithms yield the action of an arbitrary element of the group, but this action is described by a recursive algorithm rather than a closed-form formula, and the representation matrices so obtained are not in general unitary.

For the orthogonal and special orthogonal groups, straightening algorithms were developed by Lancaster–Towber (1979), and they also gave an explicit description of generators and relations for the symplectic case. A full algorithm for the symplectic groups was given by DeConcini (1979). This algorithm utilizes a completely different labelling system from ours. The only algorithm currently known for representations of an exceptional Lie group is the one for G_2 developed by Baclawski–Towber (1982).

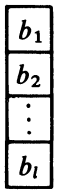
The author wishes to express his appreciation to Jacob Towber for furnishing many of the references above and for helping to check the final version of this introduction.

1. Partially ordered sets. For a more detailed discussion of the partially ordered sets (posets) considered below, see Baclawski (1982). For our needs it suffices just to define them. To each unitary or symplectic group G we define a poset called a *fundamental poset.* In addition, every element of a fundamental poset has a label attached to it. This label has two parts, each being an element of the weight lattice Λ of G (or equivalently an r-tuple of integers, where r is the rank of G). The first

part is the *irrep label* which designates the irrep to which that element of the poset belongs. The second part is the *weight* of the element.

Let $\lambda_1, \cdots, \lambda_r$ be the standard basis vectors of the weight lattice. These are the irrep labels of the fundamental irreps of G.

1.1 The special unitary groups. The fundamental poset is $\mathbf{A}(n-1) = \{(b_1 < b_2 < \cdots < b_l) | 1 \leqq l \leqq n-1 \text{ and } 1 \leqq b_1 < \cdots < b_l \leqq n\}$. The elements are usually written as columns:



The partial order is defined by: $(b_1 < b_2 < \cdots < b_l) \leqq (c_1 < \cdots < c_m)$ if and only if $l \geqq m$ and for every $i \leqq m$, we have $b_i \leqq c_i$. The irrep label of $(b_1 < b_2 < \cdots < b_l)$ is λ_l and the weight is $\sum_{j=1}^{l} (\lambda_{b_j} - \lambda_{b_j - 1})$, where we use the convention that $\lambda_0 = \lambda_n = 0$.

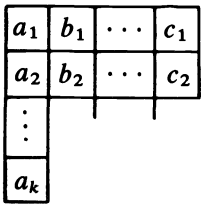
1.2. The unitary groups. The fundamental poset is $\mathbf{P}(n) = \mathbf{A}(n-1) \cup \{d, \bar{d}\}$. The partial order is defined so that $\mathbf{A}(n-1)$ is a subposet, every element of which is larger than both d and $\bar{d}$, which are not comparable to each other. The irrep label of $(b_1 < \cdots < b_l)$ is $\sum_{i=1}^{l} \lambda_i$, and the weight is $\sum_{i=1}^{l} \lambda_{b_i}$. The irrep labels of $d, \bar{d}$ are $\sum_{i=1}^{n} \lambda_i$ and $-\sum_{i=1}^{n} \lambda_i$, respectively. The weights of $d, \bar{d}$ coincide with their irrep labels. Note that the rank of $U(n)$ is n.

1.3. The symplectic groups. The fundamental poset is $\mathbf{C}(n) = \{B \in \mathbf{A}(2n-1) | B \geqq (1 < 3 < \cdots < (2n-1))\}$, with the inherited partial order. The irrep label of $(b_1 < b_2 < \cdots < b_l)$ is λ_l, while the weight is $\sum_{i=1}^{l} f(b_i)$, where

$$f(k) = \begin{cases} \lambda_j - \lambda_{j-1} & \text{if } k = 2j-1, \\ \lambda_{j-1} - \lambda_j & \text{if } k = 2j, \end{cases}$$

where we use the convention $\lambda_0 = 0$.

We will use the following notation from the theory of partially ordered sets. A *multichain* (or *standard sequence*) in a poset P is a sequence of elements $x_1, x_2, \cdots, x_k$ such that $x_1 \leqq x_2 \leqq \cdots \leqq x_k$. Note that repetitions are allowed. If P is labelled, then the label of a multichain is the sum of the labels of its elements (counting multiplicities). When P is one of the fundamental posets described above, the multichains are usually written as products, in which case they are called *standard products, standard monomials* or *Young tableaux*. For example, if $\mathbf{a} \leqq \mathbf{b} \leqq \cdots \leqq \mathbf{c}$ is a multichain in P, then the usual way to write the corresponding standard product is



The importance of the fundamental poset is that the number of multichains having a given irrep label is equal to the dimension of the corresponding irrep. Furthermore, if the weights are also considered, the same is true for the weight spaces of an irrep. These facts about the fundamental posets will be called the "labelling theorem." It was developed by a series of individuals, as described in the introduction.

2. The shape algebra. As a vector space, the shape algebra of the Lie group G, denoted $\mathbf{\Lambda}_G^+$, is simply the direct sum of all irreps of G. It is more subtle to explain how this ring is related to the fundamental poset of G and how it acquires a ring structure. We will define each shape algebra using the familiar "generators and relations" method. Roughly speaking, the generators correspond to the elements of the fundamental poset, and there is one relation for each incomparable pair of elements. As a result, the set of multichains of the fundamental poset corresponds to a basis of $\mathbf{\Lambda}_G^+$, thus associating each multichain with a specific vector in some irrep of G.

For example, $\mathbf{\Lambda}_{SU(3)}^+$ has six generators and one relation; see Fig. 1.

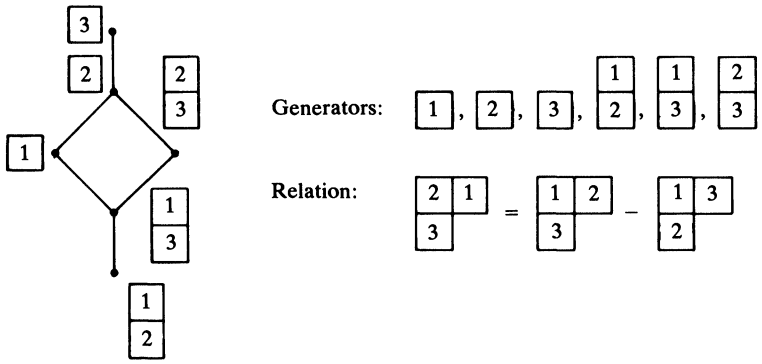


FIG. 1

In the case of $U(3)$, we have eight generators and two relations; see Fig. 2.

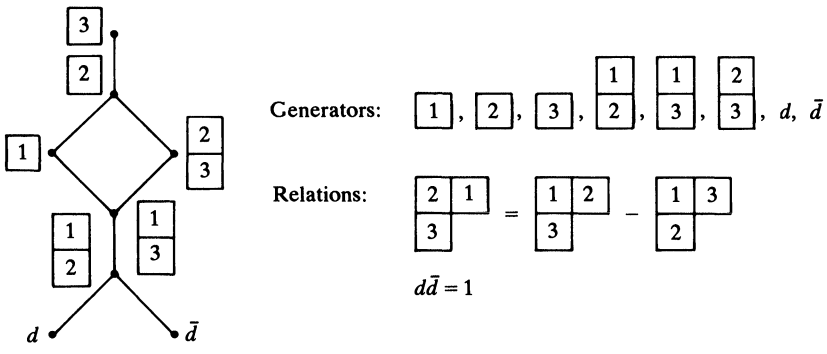
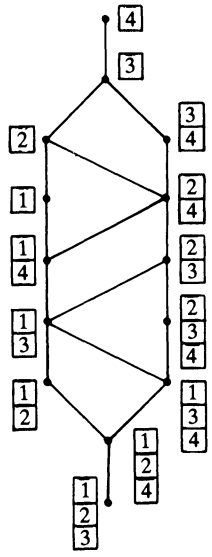


FIG. 2

The ring $\mathbf{\Lambda}_{Sp(2n,\mathbb{C})}^+$ will be constructed by starting with $\mathbf{\Lambda}_{SU(2n)}^+$ and then adjoining (or "modding out by") a linear relation for each element of $\mathbf{A}(2n-1)\backslash\mathbf{C}(n)$. For example, $Sp(4, \mathbb{C})$ has 14 generators and 10 relations; see Fig. 3.



A(3)

Relations:

$$\begin{vmatrix}1\\2\\3\end{vmatrix} = \begin{vmatrix}1\\2\\4\end{vmatrix} = \begin{vmatrix}1\\3\\4\end{vmatrix} = \begin{vmatrix}2\\3\\4\end{vmatrix} = 0$$

$$\begin{vmatrix}1\\2\end{vmatrix} = -\begin{vmatrix}3\\4\end{vmatrix}$$

$$\begin{vmatrix}1&2\\4&3\end{vmatrix} = \begin{vmatrix}1&2\\3&4\end{vmatrix} - \begin{vmatrix}1&3\\2&4\end{vmatrix}$$

$$\begin{vmatrix}2&1\\3\end{vmatrix} = \begin{vmatrix}1&2\\3\end{vmatrix} - \begin{vmatrix}1&3\\2\end{vmatrix}$$

$$\begin{vmatrix}2&1\\4\end{vmatrix} = \begin{vmatrix}1&2\\4\end{vmatrix} - \begin{vmatrix}1&4\\2\end{vmatrix}$$

$$\begin{vmatrix}3&1\\4\end{vmatrix} = \begin{vmatrix}1&3\\4\end{vmatrix} - \begin{vmatrix}1&4\\3\end{vmatrix}$$

$$\begin{vmatrix}3&2\\4\end{vmatrix} = \begin{vmatrix}2&3\\4\end{vmatrix} - \begin{vmatrix}2&4\\3\end{vmatrix}$$

Shape algebra $\Lambda^{+}_{Sp(4)}$

FIG. 3

We now give the rigorous definitions of the shape algebras $\Lambda^{+}_{SU(n)}$ and $\Lambda^{+}_{U(n)}$ of the unitary groups. Although this is a well-known classical case, we include a discussion for the sake of completeness.

2.1. The special unitary groups. The fundamental poset of this group is $\mathbf{A}(n-1)$. We regard the elements of $\mathbf{A}(n-1)$ as indeterminates or generators of a ring whose relations are called the *shuffle relations*. To define these it is useful to define a total order on the elements of $\mathbf{A}(n-1)$ called the *lexicographic order*. Given $\mathbf{b} = (b_1 < \cdots < b_k)$, $\mathbf{c} = (c_1 < \cdots < c_l) \in \mathbf{A}(n-1)$, we say that $\mathbf{b}$ (*strictly*) *precedes* $\mathbf{c}$ if

(1) $k \geqq l$,

(2) if $k = l$, then for some i, $b_1 = c_1, \cdots, b_{i-1} = c_{i-1}$ and $b_i < c_i$.

Now let $\mathbf{b}$ and $\mathbf{c}$ be any pair of incomparable elements of $\mathbf{A}(n-1)$. We may assume that $\mathbf{b}$ precedes $\mathbf{c}$. Now since $\mathbf{b}$ and $\mathbf{c}$ are not comparable, there is a first index j, called the *violation index* of $\mathbf{bc}$, such that $b_j \not\leqq c_j$. We then have that $c_1 < c_2 < \cdots < c_j < b_j < b_{j+1} < \cdots < b_k$. Let $\mathcal{SH}(\mathbf{bc})$ be the set of all permutations τ of $\{c_1, \cdots, c_j, b_j, \cdots, b_k\}$ such that $\tau(c_1) < \cdots < \tau(c_j)$ and $\tau(b_j) < \cdots < \tau(b_k)$. For $\tau \in \mathcal{SH}(\mathbf{bc})$, we define

$$\tau(\mathbf{bc}) = \text{incr}\,(b_1, \cdots, b_{j-1}, \tau(b_j), \cdots, \tau(b_k)) \cdot \text{incr}\,(\tau(c_1), \cdots, \tau(c_j), c_{j+1}, \cdots, c_l),$$

where if $d_1, \cdots, d_l$ is a sequence of integers, then we write

$$\text{incr}\,(d_1, \cdots, d_l) = \begin{cases} 0 & \text{if } d_i = d_j \quad \text{for some } i \neq j, \\ \text{sgn}\,(\sigma)(\sigma(d_1) < \cdots < \sigma(d_l)) & \text{otherwise;} \end{cases}$$

where σ is the permutation required to put $d_1, \cdots, d_l$ in increasing order and sgn (σ) is the sign of σ.

The *shuffle relation* of $\mathbf{bc}$ is then given by:

$$\sum_{\tau \in \mathcal{SH}(\mathbf{bc})} \text{sgn}\,(\tau)\tau(\mathbf{bc}) = 0.$$

If we write e for the identity permutation, then $e(\mathbf{bc})=\mathbf{bc}$; hence, we may also write the shuffle relation in this form:

$$\mathbf{bc} = -\sum_{\substack{\tau\in\mathscr{SH}(\mathbf{bc})\\ \tau\neq e}} \operatorname{sgn}(\tau)\tau(\mathbf{bc}).$$

Now the products $\tau(\mathbf{bc})$ for $\tau\neq e$ need not be standard; however, the violation index of $\tau(\mathbf{bc})$ is strictly larger than j for every $\tau\neq e$. So if we recursively substitute the shuffle relation for those products $\tau(\mathbf{bc})$ that are nonstandard, we eventually obtain a relation of the form

$$\mathbf{bc}=\sum_i r_i\mathbf{b}_i\mathbf{c}_i,$$

where each product $\mathbf{b}_i\mathbf{c}_i$ is standard and each r_i is an integer.

From this one can show that the standard products span the ring $\mathbf{\Lambda}^+_{SU(n)}$. That the standard products are linearly independent is harder to show and is essentially due to Hodge (1941) (in characteristic zero) and Doubilet et al. (1974) (in general). See also DeConcini et al. (1980).

We now examine how $SU(n)$ acts on $\mathbf{\Lambda}^+_{SU(n)}$. It suffices to describe how $Gl(n,\mathbb{C})$ acts on the generators of $\mathbf{\Lambda}^+_{SU(n)}$ and to show that $Gl(n,\mathbb{C})$ maps the ideal generated by the shuffle relations into itself. The generators of $\mathbf{\Lambda}^+_{SU(n)}$ or equivalently the elements of $\mathbf{A}(n-1)$ should be thought of as the basis vectors of the fundamental irreps: $V(\lambda_1), V(\lambda_2), \cdots, V(\lambda_{n-1})$. In more concrete terms, $V(\lambda_1)$ is the vector space $\mathbb{C}^n$ upon which $Gl(n,\mathbb{C})$ acts in the natural way. The basis vectors are the one-element sequences $\boxed{1}, \cdots, \boxed{n}$ in $\mathbf{A}(n-1)$. The other fundamental irreps are the exterior powers of $\mathbb{C}^n$, the basis of $V(\lambda_j)=\Lambda^j\mathbb{C}^n$ being all sequences $(a_1<\cdots<a_j)\in\mathbf{A}(n-1)$ having exactly j elements. A sequence $(a_1<\cdots<a_j)$ should be thought of as the wedge product $\boxed{a_1}\wedge\cdots\wedge\boxed{a_j}$ of these j vectors from $\mathbb{C}^n$.

It is an exercise in linear algebra to show that for any $\sigma\in Gl(n,\mathbb{C})$, the action of σ on a shuffle relation is a linear combination of shuffle relations. In fact, the coefficients will be determinants of certain submatrices of σ. This is shown in Doubilet et al. (1974) and DeConcini et al. (1980).

Thus $\mathbf{\Lambda}^+_{SU(n)}$ is a well-defined ring with an action of $SU(n)$. We now consider how this representation decomposes as a direct sum of irreps. For this we need the labels we attached to the elements and the multichains of $\mathbf{A}(n-1)$. Now the multichains form a basis of $\mathbf{\Lambda}^+_{SU(n)}$ which behaves extremely well with respect to the action of $SU(n)$. For a given irrep label, λ, the space $\mathbf{\Lambda}^+_\lambda$ of all linear combinations of standard monomials having this irrep label, is an $SU(n)$-submodule of $\mathbf{\Lambda}^+_{SU(n)}$, since the action of $SU(n)$, as well as the shuffle relations, preserve irrep labels. In ring theory one calls $\mathbf{\Lambda}^+_\lambda$ the *homogeneous component* of $\mathbf{\Lambda}^+_{SU(n)}$ having *multidegree* λ. To determine how the representation $\mathbf{\Lambda}^+_\lambda$ breaks up as a sum of irreps, it suffices to compute its *formal character*: the formal power series

$$F_\lambda(\alpha)=\sum_\mu \dim(\mathbf{\Lambda}^+_{\lambda,\mu})\alpha_1^{\mu_1}\alpha_2^{\mu_2}\cdots\alpha_{n-1}^{\mu_{n-1}},$$

where $\mathbf{\Lambda}^+_{\lambda,\mu}$ is the weight space of $\mathbf{\Lambda}^+_\lambda$ corresponding to weight μ. Now $\mathbf{\Lambda}^+_{\lambda,\mu}$ is precisely the subspace spanned by standard monomials whose irrep label is λ and whose weight label is μ. By the labelling theorem, $F_\lambda(\alpha)$ coincides with the formal character of the irrep $V(\lambda)$. Thus $\mathbf{\Lambda}^+_\lambda\cong V(\lambda)$, and it follows that $\mathbf{\Lambda}^+_{SU(n)}$ contains every irrep of $SU(n)$ exactly once.

2.2. The unitary groups. As a ring we construct $\Lambda^+_{U(n)}$ from $\Lambda^+_{SU(n)}$ by adjoining two new generators and one new relation:

$$\Lambda^+_{U(n)} = \Lambda^+_{SU(n)}[d, \bar{d}]/(d\bar{d} = 1).$$

The action of $Gl(n, \mathbb{C})$ on $\Lambda^+_{U(n)}$ is obtained from that on $\Lambda^+_{SU(n)}$ as follows: for $\sigma \in Gl(n, \mathbb{C})$, define

$$\sigma(d) = \det(\sigma)d \quad \text{and} \quad \sigma(\bar{d}) = \det(\sigma)^{-1}\bar{d}.$$

One should think of d as the sequence $(1<2<\cdots<n)$ or, equivalently, the wedge product $\boxed{1} \wedge \boxed{2} \wedge \cdots \wedge \boxed{n}$.

3. The symplectic groups. The construction of the shape algebra for the symplectic groups is more involved than the construction for the unitary groups. The idea is to begin with $\Lambda^+_{SU(2n)}$ and to adjoin linear relations to get $\Lambda^+_{Sp(2n)}$. This is not the only approach one can use, but it is convenient and less cumbersome than the other methods mentioned in the introduction.

We now outline the procedure we will follow in our construction. Let $V_1, \cdots, V_{2n-1}$ be the fundamental irreps of $Sl(2n, \mathbb{C})$, i.e., $V_j = \Lambda^j \mathbb{C}^{2n}$; and let $V'_1, \cdots, V'_n$ be the fundamental irreps of $Sp(2n, \mathbb{C})$. It is well known that when $Sl(2n, \mathbb{C})$ is restricted to $Sp(2n, \mathbb{C})$, then for $j = 1, 2, \cdots, n$, V_j contains V'_j as a submodule, i.e., $V_j = V'_j \oplus V''_j$ for a suitable module V''_j over $Sp(2n, \mathbb{C})$. Thus we have $V'_j \cong V_j / V''_j$, and so if we can compute V''_j, then we have a model for V'_j. Since V''_j is an $Sp(2n, \mathbb{C})$-submodule of V_j, it follows that the ideal generated by V''_j in $\Lambda^+_{SU(2n)}$ is an $Sp(2n, \mathbb{C})$-submodule also. Let I be the ideal generated by V''_j, $1 \leqq j \leqq n$, and by V_j, $n+1 \leqq j \leqq 2n-1$. Then $\Lambda^+_{SU(2n)}/I$ contains every irrep of $Sp(2n, \mathbb{C})$ at least once. By computing the dimension of each homogeneous part of $\Lambda^+_{SU(2n)}/I$ and comparing it to the known dimensions of the irreps of $Sp(2n, \mathbb{C})$, we finally conclude that $\Lambda^+_{SU(2n)}/I$ has each irrep exactly once and hence that $\Lambda^+_{Sp(2n)} \cong \Lambda^+_{SU(2n)}/I$.

The first step, then, is to compute V''_j, and we do this by using the Casimir operator. Let C be the symplectic Casimir operator. Let ρ be the sum $\sum_{j=1}^n \lambda_j$ of the fundamental irrep labels in the weight lattice Λ of $Sp(2n, \mathbb{C})$. We give Λ the usual inner product (induced by the Killing form). Then the Casimir operator C simply multiplies each vector in V'_j by $(\lambda_j + \rho, \lambda_j + \rho)$. Thus the image of the operator $C - (\lambda_j + \rho, \lambda_j + \rho)$ on V_j is a subspace V'''_j of V''_j (which we will eventually show is the same as V''_j).

To compute $C - (\lambda_j + \rho, \lambda_j + \rho)$ we need some auxiliary notation. Let $V = \bigoplus \sum_{j=0}^{2n} V_j$, where $V_0 = V_{2n} = \mathbb{C}$. Then V_k is a vector space whose basis is $\{(a_1 < \cdots < a_k) | a_i \leqq 2n\}$. The basis vectors of V will be thought of both as sequences and as subsets of $[2n] = \{1, 2, \cdots, 2n\}$. For a subset $S \subseteq [n]$ and a function $\varepsilon : S \to \{0, 1\}$, define $V(S, \varepsilon)$ to be the subspace of V consisting of linear combinations of

$$\{A = (a_1 < \cdots < a_k) \subseteq [2n] \Big| \text{if } j \notin S, \text{ then } |\{2j-1, 2j\} \cap A| \neq 1 \text{ and if } j \in S, \text{ then } \{2j-1, 2j\} \cap A = \{2j - \varepsilon\}\}.$$

For a given $A \subseteq [2n]$, we see that $A \in V(S, \varepsilon)$ for some ε if and only if $S = \{j \,|\, |\{2j-1, 2j\} \cap A| = 1\}$. It is easy to see that $V = \bigoplus \sum_{S,\varepsilon} V(S, \varepsilon) = \bigoplus \sum_{S,\varepsilon,i} V_i(S, \varepsilon)$, where $V_i(S, \varepsilon) = V(S, \varepsilon) \cap V_i$.

For a given choice of S and ε, the basis vectors of $V(S, \varepsilon)$ are distinguished from one another uniquely by

$$D(\mathring{A}) = \{j \in [n] \Big| |\{2j-1, 2j\} \cap A| = 2\}.$$

If $A \in V_i(S, \varepsilon)$, then $|S| + 2|D(A)| = i$. Now by definition, we have that $D(A) \subseteq [n] \backslash S$, and conversely it is obvious that every subset of $[n] \backslash S$ has a corresponding basis vector A of $V(S, \varepsilon)$. Henceforth write $m = n - |S|$. Let $\phi : [n] \backslash S \to [m]$ be the order-preserving bijection. Then $\phi \circ D$ maps the basis vectors of $V(S, \varepsilon)$ bijectively onto the Boolean algebra $\mathcal{B}(m)$. Let $W(m)$ be the vector space on $\mathcal{B}(m)$ as a basis. Extending $\phi \circ D$ linearly, we have a linear isomorphism $\psi : V(S, \varepsilon) \tilde{\to} W(m)$. Write $W_l(m)$ for the subspace of $W(m)$ spanned by l-element subsets of $[m]$. Then we have

$$V_{|S|+2l}(S, \varepsilon) \cong W_l(m).$$

We now define operators $X, Y : W(m) \to W(m)$ as follows:

$$X(L) = \sum_{p \notin L} L \cup \{p\}, \qquad Y(M) = \sum_{p \in M} M \backslash \{p\}.$$

These operators and their brackets define an action of the Lie algebra $sl(2, \mathbb{C})$ on $W(m)$; see Stanley (1982) and Proctor (1979). We will use the fact that for $l \leqq m/2$, $X : W_{l-1}(m) \to W_l(m)$ is injective and $Y : W_l(m) \to W_{l-1}(m)$ is surjective (although strictly speaking this fact follows from Theorem 4). Let $Z : W(m) \to W(m)$ be the composition $X \circ Y$. By a "brute force" computation one can verify that:

LEMMA 1. *For every $l \leqq m/2$, this diagram commutes*:

$$\begin{array}{ccc} & V_j(S, \varepsilon) & \xrightarrow{\tilde{\psi}} W_l(m) \\ C - (\lambda_j + \rho, \lambda_j + \rho) & \downarrow & \downarrow Z \\ & V_j(S, \varepsilon) & \xrightarrow{\tilde{\psi}} W_l(m), \end{array}$$

where $j = |S| + 2l$.

Note that the condition $l \leqq m/2$ is equivalent to $j \leqq n$, which is exactly the range we are interested in. As noted earlier, in this range Y is surjective, so the image of Z coincides with the image of X. Moreover, in this range X is injective. Thus we have a relatively concrete description of V_j''': it is the direct sum

$$\bigoplus \sum_{\substack{S, \varepsilon, l \\ |S|+2l=j}} \psi_{S,\varepsilon}^{-1} X(W_l(n - |S|)).$$

We next seek a basis for the quotient V_j / V_j''' and an algorithm for expressing an arbitrary vector $v + V_j'''$, for $v \in V_j$, in terms of this basis. By Lemma 1, it suffices to find a basis and algorithm for the quotients $W_l(m)/X(W_{l-1}(m))$, where $l \leqq m/2$. A subset $P \subseteq [m]$ will be called a *Yamanouchi* subset if $P = (p_1 < p_2 < \cdots < p_k)$ satisfies $p_i \geqq 2i$ for every i. Let $\mathcal{Y}am_l$ be the Yamanouchi subsets of cardinality l. A beautiful construction of Vo (1981) gives an injective map $\mathcal{B}_{l-1}(m) \to \mathcal{B}_l(m)$, if $l \leqq m/2$, whose image consists of all non-Yamanouchi subsets of cardinality l. Thus $|\mathcal{Y}am_l| = \binom{m}{l} - \binom{m}{l-1}$.

The Yamanouchi condition arose because of the following:

THEOREM 2. *Let $A = (a_1 < \cdots < a_k) \in \mathbf{A}(2n-1)$ and define S, ε so that $A \in V_k(S, \varepsilon)$. Then $A \in \mathbf{C}(n)$ if and only if $\psi_{S,\varepsilon}(A)$ is a Yamanouchi subset.*

Proof. Let $m = n - |S|$ and $l = (k - |S|)/2$ so that $\psi_{S,\varepsilon}(A) \in W_l(m)$. Clearly $A \in \mathbf{C}(n)$ if and only if $A \cap [2s] \in \mathbf{C}(s)$ for every s, and similarly for the Yamanouchi condition. By induction we assume that the result is true for n replaced by $n-1$.

Suppose that $A \in \mathbf{C}(n)$. Then $\psi(A \cap [2n-2]) = (p_1 < \cdots < p_e)$ satisfies $p_i \geqq 2i$ for every i. Now e must be either l or $l-1$. If $e = l$, we are done. So we may assume that

$e = l-1$. In this case $p_l = m$. Suppose that $p_l < 2l$. Then since $p_{l-1} \geqq 2l-2$ and $p_{l-1} < p_l$, we have that $p_l = 2l-1$ and hence that $m = 2l-1$. This, in turn, implies that $k = |A| = n+1$. Now $p_l = 2l-1$ means that the last two elements of A, a_n and a_{n+1}, are not in $[2n-2]$. Thus $a_n = 2n-1$ and $a_{n+1} = 2n$, but this contradicts the assumption that $A \in \mathbf{C}(n)$. Hence $\psi(A)$ is Yamanouchi as desired.

Conversely, suppose that $\psi(A)$ is Yamanouchi. By induction we may assume that $A \cap [2n-2] \in \mathbf{C}(n-1)$, so we need only show that $a_k \geqq 2k-1$ and that $a_{k-1} \geqq 2k-3$. There are three possible cases:

Case 1. $a_k \leqq 2n-2$. Nothing to show, by induction.

Case 2. $a_k > 2n-2$, $a_{k-1} \leqq 2n-2$. Suppose that $a_k < 2k-1$. By induction, $a_{k-1} \geqq 2k-3$, so it follows that $a_{k-1} = 2k-3$ and that $a_k = 2k-2$. Since $a_k \geqq 2n-1$ and a_k is even, it follows that $a_k = 2n$. Thus $k = n+1$, which immediately contradicts the fact that $a_{k-1} \leqq 2n-2$.

Case 3. $a_{k-1} > 2n-2$. This immediately implies that $a_k = 2n$ and that $a_{k-1} = 2n-1$. We must show that $a_{k-1} \geqq 2k-3$ and that $a_k \geqq 2k-1$. If either of these inequalities fails, then the second must fail. Suppose it does, i.e., suppose $a_k < 2k-1$. In this case $p_l = m$, so by the hypothesis that $\psi(A)$ is Yamanouchi, we have that $m \geqq 2l$. Thus $k = n-m+2l \leqq n$. However, this in turn implies that $2n = a_k < 2k-1 \leqq 2n-1$, a contradiction. The result then follows. Q.E.D.

Let Yam_l be the subspace of $W_l(m)$ spanned by $\mathscr{Y}am_l$. Then $\dim X(W_{l-1}(m)) + \dim Yam_l = \dim W_l(m)$. Accordingly, if we can show that $X(W_{l-1}(m)) + Yam_l$ spans $W_l(m)$, then it will follow that $Yam_l \cong W_l(m)/X(W_{l-1}(m))$. In fact, we will describe a recursive algorithm for expressing any basis vector $Q \in \mathscr{B}_l(m)$ as a linear combination of Yamanouchi subsets, modulo $X(W_{l-1}(m))$.

We first need

LEMMA 3. *The matrix* $(\chi(L \subseteq M))_{|L|=l-1,|M|=l}$, *where* $L, M \subseteq [2l-1]$, *is a square matrix whose inverse is given by* $(f_l(|L \cap M|))_{|M|=l,|L|=l-1}$, *where*

$$f_l(j) = \frac{(-1)^{l-j-1}}{l\binom{l-1}{j}} \quad \text{and} \quad \chi(L \subseteq M) = \begin{cases} 1 & \text{if } L \subseteq M, \\ 0 & \text{if } L \not\subseteq M. \end{cases}$$

We leave the proof as an exercise.

Now for $Q = (q_1 < \cdots < q_l) \in \mathscr{B}(m)$, define $h(Q) = \max\{j | j = 0 \text{ or } q_j < 2j\}$, and write $Q_1 = (q_1 < \cdots < q_{h(Q)})$ and $Q_2 = Q \backslash Q_1$. Note that Q is Yamanouchi if and only if $h(Q) = 0$, so $h(Q)$ is a measure of the non-Yamanouchiness of Q. Our key formula is

THEOREM 4. *Let* $Q \in \mathscr{B}_l(m)$ *for some* $l \leqq m/2$, *and let* $h = h(Q)$. *Then*

$$X\Bigg(\sum_{\substack{M \subseteq [2h-1] \\ |M|=h-1}} f_h(|Q_1 \cap M|) M \cup Q_2\Bigg) = Q + \sum_{\substack{M \subseteq [2h-1] \\ |M|=h-1}} f_h(|Q_1 \cap M|) \sum_{\substack{t=2h \\ t \notin Q_2}}^{m} M \cup \{t\} \cup Q_2.$$

Proof. We begin with

$$(*) \qquad X\Bigg(\sum_{\substack{M \subseteq [2h-1] \\ |M|=h-1}} f_h(|Q_1 \cap M|) M \cup Q_2\Bigg) = \sum_{\substack{M \subseteq [2h-1] \\ |M|=h-1}} f_h(|Q_1 \cap M|) \sum_{t \notin M \cup Q_2} M \cup \{t\} \cup Q_2.$$

We break the final summation into $\sum_{t\leqq 2h-1,t\notin M}+\sum_{t\geqq 2h,t\notin Q_2}$. Now $\sum_{t\leqq 2h-1,t\notin M} M\cup\{t\}\cup Q_2=\sum_{M\subset N\subset[2h-1],|N|=h} N\cup Q_2$, so the first term in (*) may be written as follows (by Lemma 3):

$$\sum_{\substack{N\subseteq[2h-1]\\|N|=h}}\left(\sum_{\substack{M\subseteq[2h-1]\\|M|=h-1}} f_h(|Q_1\cap M|)\chi(M\subset N)N\cup Q_2\right)$$

$$=\sum_{\substack{N\subseteq[2h-1]\\|N|=h}}\chi(Q_1=N)N\cup Q_2=Q_1\cup Q_2=Q.$$

Since the second term in (*) is the second term of the desired formula, we are done. Q.E.D.

It follows that

$$Q\equiv-\sum_{\substack{M\subseteq[2h-1]\\|M|=h-1}} f_h(|Q_1\cap M|)\sum_{\substack{t=2h\\t\notin Q_2}}^{m} M\cup\{t\}\cup Q_2 \text{ (modulo } X(W_{l-1}(m))).$$

Since $h(M\cup\{t\}\cup Q_2)<h$ whenever $t\geqq 2h$, this formula gives a recursive algorithm for expressing any $Q\in\mathcal{B}_l(m)$ in terms of Yam_l, modulo $X(W_{l-1}(m))$. We have thus shown:

COROLLARY 5. *For any* $l\leqq m/2$, $Yam_l\cong W_l(m)/X(W_{l-1}(m))$.

Combining Lemma 1, Theorem 2 and Corollary 5, we have that $\mathbf{C}(n)$ is a basis of $\bigoplus\sum_{j=1}^{n} V_j/V_j'''$. However, by the labelling theorem, $|\mathbf{C}(n)|=\dim(\bigoplus\sum_{j=1}^{n} V_j')=\dim(\bigoplus\sum_{j=1}^{n} V_j/V_j'')$. Since $V_j'''\subseteq V_j''$, we conclude that $V_j'''=V_j''$, for every j, and that $\mathbf{C}(n)$ is a basis of $\bigoplus\sum_{j=1}^{n} V_j/V_j''\cong\bigoplus\sum_{j=1}^{n} V_j'$.

It is perhaps useful to give an intuitive picture of the formula in Theorem 4. To do so we borrow some terminology from the game of bridge. We regard an element $\mathbf{b}$ of $\mathbf{A}(2n-1)$ as a "hand" A from the "deck" $[2n]$, whose elements form n "suits," the first suit being $\{1,2\}$, the second $\{3,4\}$, etc. Those suits for which A possesses exactly one element, the "singletons," are the set S; the set of "doubletons" is $D(A)$, while all other elements of $[n]$ are the "voids" of A. The number of doubletons and voids is denoted m. If $\mathbf{b}$ fails to be in $\mathbf{C}(n)$, it does so because of the first h doubletons, where $h=\max\{j|j=0 \text{ or } q_j<2j\}$ and where q_j is the number of doubleton or void suits up to the jth doubleton suit in A. Theorem 4 then dictates how to "shuffle" doubletons into voids to yield elements of $\mathbf{A}(2n-1)$ closer to being in $\mathbf{C}(n)$.

For example, if $n=3$ and $\mathbf{b}=(1<2<3)$, then $S=\{2\}$, $D=\{1\}$ and there is just one void. Theorem 4 yields that $(1<2<3)=-(3<5<6)$. Similarly if $\mathbf{b}=(1<2)$, then $S=\phi$, $D=\{1\}$ and there are now two voids. Theorem 4 then yields that $(1<2)=-(3<4)-(5<6)$.

We conjecture, based on some computational evidence, that the expansion of an element of $\mathbf{A}(2n-1)\backslash\mathbf{C}(n)$ as a linear combination of elements of $\mathbf{C}(n)$ involves only integral coefficients, and moreover, that these coefficients have a combinatorial interpretation. This would speed up our algorithm by making this part nonrecursive.

We now consider the whole ring $\mathbf{\Lambda}^+_{SU(2n)}/I$. Using the Casimir operator method once again, it is easy to see that $\mathbf{\Lambda}^+_{Sp(2n)}$ is a quotient of $\mathbf{\Lambda}^+_{SU(2n)}/I$. Fix an irrep label λ and let $\mathbf{\Lambda}^+_\lambda, \mathbf{\Lambda}'_\lambda$ be the homogeneous components of the former and latter rings respectively. By definition, $\mathbf{\Lambda}^+_\lambda$ is the irrep $V(\lambda)$ of $Sp(2n,\mathbb{C})$. By the labelling theorem once again, $\dim(V(\lambda))$ is the number of multichains of $\mathbf{C}(n)$ having irrep label λ. But we have just shown (and we discuss more thoroughly in the next section) that these

multichains span $\mathbf{\Lambda}'_\lambda$. Hence $\dim \mathbf{\Lambda}'_\lambda \leqq \dim \mathbf{\Lambda}^+_\lambda$, while $\mathbf{\Lambda}^+_\lambda$ is a quotient of $\mathbf{\Lambda}'_\lambda$. Therefore $\mathbf{\Lambda}^+_{Sp(2n)} = \mathbf{\Lambda}^+_{SU(2n)}/I$ and our description of $\mathbf{\Lambda}^+_{Sp(2n)}$ is complete.

4. Algorithms. We now collect some of the material in the previous sections and present it in more algorithmic form. The problem is to compute the action of an element σ of the given Lie group G on a multichain $(\mathbf{x}_1 \leqq \cdots \leqq \mathbf{x}_k)$ of the fundamental poset P of G.

We first discuss the case of a unitary group representation. To compute $\sigma(\mathbf{x}_1 \cdots \mathbf{x}_k)$, we apply σ to each factor and expand the product $\sigma(\mathbf{x}_1)\sigma(\mathbf{x}_2)\cdots\sigma(\mathbf{x}_k)$ as a linear combination of monomials. The monomials that occur will not, in general, be standard. Suppose that $\mathbf{y}_1\mathbf{y}_2\cdots\mathbf{y}_k$ is such a monomial. Then at least one pair of factors, say $\mathbf{y}_i$ and $\mathbf{y}_j$, will not be comparable. To find this pair efficiently, one arranges the factors according to the lexicographic criterion discussed in § 2.1. When this is done, there will be at least one adjacent pair of incomparable factors.

Having found an incomparable pair of factors $\mathbf{y}_i, \mathbf{y}_j$, we utilize the corresponding shuffle relation to replace the product $\mathbf{y}_i\mathbf{y}_j$ by a linear combination $\sum_l r_l\mathbf{b}_l\mathbf{c}_l$. Substitute this into the monomial $\mathbf{y}_1\mathbf{y}_2\cdots\mathbf{y}_k$ and expand to yield a linear combination whose terms are all "closer" to being standard. If we apply this technique recursively, the procedure eventually terminates with a linear combination of standard monomials. The whole algorithm is called a *lexicographic straightening algorithm* because each substitution of a shuffle relation yields a linear combination of monomials each of which strictly precedes the original monomial with respect to a lexicographic order on monomials, thereby ensuring that the algorithm eventually terminates.

In the case of a symplectic representation, the algorithm has an additional step. As before, we expand the product $\sigma(\mathbf{x}_1)\sigma(\mathbf{x}_2)\cdots\sigma(\mathbf{x}_k)$ as a linear combination of monomials. Let $\mathbf{y}_1\mathbf{y}_2\cdots\mathbf{y}_k$ be one such monomial. It is possible that some $\mathbf{y}_i$ is not in $\mathbf{C}(n)$. If this happens, replace one such $\mathbf{y}_i$ by the expression given by Theorem 4, and expand. Do this recursively until only elements of $\mathbf{C}(n)$ occur as factors. One then proceeds as in the unitary case, except that if a factor from $\mathbf{A}(2n-1)\backslash\mathbf{C}(n)$ is ever introduced, the procedure above must again be applied. Ultimately, the result will be a linear combination of standard monomials whose factors are in $\mathbf{C}(n)$. Once again we get a lexicographic straightening law, but the lexicographic order on monomials used here is *not* the one for $\mathbf{A}(2n-1)$ mentioned above and used for the unitary groups. In fact, in this case larger elements of $\mathbf{A}(2n-1)$ are regarded as being "earlier" than smaller ones, while elements of $\mathbf{C}(n)$ precede all other elements of $\mathbf{A}(2n-1)$ having the same irrep label.

Although the algorithm described above seems quite complicated, it has some desirable features:

1. The order in which the quadratic and linear substitutions are made does not affect the final answer.

2. Each substitution produces a linear combination of terms, each of which strictly precedes the original term with respect to a suitable lexicographic order on monomials.

3. Each substitution preserves the weight of the monomial. Thus the speed of the algorithm depends on:

(a) the complexity of σ, or more precisely, the number of different weights occurring in the product $\sigma(\mathbf{x}_1)\sigma(\mathbf{x}_2)\cdots\sigma(\mathbf{x}_k)$; and

(b) the multiplicities of the weights occurring in $\sigma(\mathbf{x}_1)\sigma(\mathbf{x}_2)\cdots\sigma(\mathbf{x}_k)$.

Since multiplicities of weights do not grow as rapidly as the dimensions of irreps, if σ is not too complicated (e.g., a "raising" or "lowering" operator), then this algorithm would be quite efficient and fast.

We end with an example. Let $\sigma \in Sp(6, \mathbb{C})$ be

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Consider the standard monomial (written in tableau notation)

$$\begin{array}{|c|c|}\hline 3 & 5 \\ \hline 5 & 6 \\ \hline 6 \\ \cline{1-1} \end{array} \in V_{(-1,1,0)}(0, 1, 1),$$

where this means that this vector has weight $(-1, 1, 0)$ and belongs to the irrep whose highest weight is $(0, 1, 1) = \lambda_2 + \lambda_3$. Now $\boxed{1}, \cdots, \boxed{6}$ are the standard basis vectors of $\mathbb{C}^6$, so $\sigma\boxed{3} = \boxed{3}$, $\sigma\boxed{5} = \boxed{1}$ and $\sigma\boxed{6} = \boxed{2}$. Similarly,

$$\sigma \begin{array}{|c|}\hline 5 \\ \hline 6 \\ \cline{1-1} \end{array} = \begin{array}{|c|}\hline 1 \\ \hline 2 \\ \cline{1-1} \end{array} \quad \text{and} \quad \sigma \begin{array}{|c|}\hline 3 \\ \hline 5 \\ \hline 6 \\ \cline{1-1} \end{array} = \begin{array}{|c|}\hline 3 \\ \hline 1 \\ \hline 2 \\ \cline{1-1} \end{array} = \begin{array}{|c|}\hline 1 \\ \hline 2 \\ \hline 3 \\ \cline{1-1} \end{array}.$$

However,

$$\begin{array}{|c|}\hline 1 \\ \hline 2 \\ \cline{1-1} \end{array}, \quad \begin{array}{|c|}\hline 1 \\ \hline 2 \\ \hline 3 \\ \cline{1-1} \end{array} \notin \mathbf{C}(3),$$

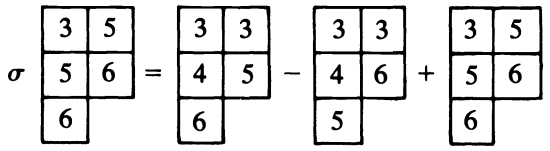
so we must apply the formula in Theorem 4. This yields

$$\begin{array}{|c|}\hline 1 \\ \hline 2 \\ \cline{1-1} \end{array} = - \begin{array}{|c|}\hline 3 \\ \hline 4 \\ \cline{1-1} \end{array} - \begin{array}{|c|}\hline 5 \\ \hline 6 \\ \cline{1-1} \end{array} \quad \text{and} \quad \begin{array}{|c|}\hline 1 \\ \hline 2 \\ \hline 3 \\ \cline{1-1} \end{array} = - \begin{array}{|c|}\hline 3 \\ \hline 5 \\ \hline 6 \\ \cline{1-1} \end{array}.$$

Thus

$$\sigma \begin{array}{|c|c|}\hline 3 & 5 \\ \hline 5 & 6 \\ \hline 6 \\ \cline{1-1} \end{array} = \begin{array}{|c|c|}\hline 3 & 3 \\ \hline 5 & 4 \\ \hline 6 \\ \cline{1-1} \end{array} + \begin{array}{|c|c|}\hline 3 & 5 \\ \hline 5 & 6 \\ \hline 6 \\ \cline{1-1} \end{array}.$$

The first of these is not standard, so we apply the shuffle relation, the violation index in this case being 2. Hence



REFERENCES

[1] K. BACLAWSKI, *Rings with lexicographic straightening law*, Adv. in Math., 39 (1981), pp. 185–213.

[2] ———, *Character-generators for unitary and symplectic groups*, submitted to J. Math. Phys., 1982.

[3] K. BACLAWSKI AND J. TOWBER, *The shape-algebra and standard bases for G_2*, submitted to Amer. J. Math., 1982.

[4] G. BAIRD AND L. BIEDENHARN, *On the representations of the semi-simple Lie groups. II*, J. Math. Phys., 4 (1963), pp. 1449–1466.

[5] J. BELINFANTE AND B. KOLMAN, *A Survey of Lie Groups and Lie Algebras with Applications and Computational Methods*, Society for Industrial and Applied Mathematics, Philadelphia 1972.

[6] C. DECONCINI, *Symplectic standard tableaux*, Adv. in Math., 34 (1979), pp. 1-27.

[7] C. DECONCINI, D. EISENBUD AND C. PROCESI, *Young diagrams and determinantal varieties*, Invent. Math., 56 (1980), pp. 129–165.

[8] ———, *Hodge algebras*, preprint, Brandeis University, Waltham, MA, 1981.

[9] C. DECONCINI AND C. PROCESI, *A characteristic free approach to invariant theory*, Adv. in Math., 21 (1976), pp. 330–354.

[10] P. DOUBILET, G.-C. ROTA AND J. STEIN, *On the foundations of combinatorial theory IX: Combinatorial methods in invariant theory*, Stud. Appl. Math., 8 (1974), pp. 185–216.

[11] H. GARNIR, *Théorie de la représentation linéaire des groupes symétriques*, Mem. Soc. Royale Sci. Liège, 4 (10) (1950), pp. 5–100.

[12] I. GELFAND AND M. ZETLIN, *Finite-dimensional representations of the group of unimodular matrices*; *Finite-dimensional representations of groups of orthogonal matrices*, Dokl. Akad. Nauk SSSR, 71 (1950), 825–828; 1017–1020 (In Russian).

[13] M. HAMERMESH, *Group Theory and Its Application to Physical Problems*, Addison-Wesley, Reading, MA, 1962.

[14] G. HEGERFELDT, *Branching rules for the symplectic groups*, J. Math. Phys., 8 (1967), pp. 1195–1196.

[15] W. HODGE, *The base for algebraic varieties of given dimension on a Grassmannian variety*, J. London Math. Soc., 16 (1941), pp. 245–255.

[16] ———, *A note on k-connexes*, Proc. Cambridge Philos. Soc., 38 (1942), pp. 129–143.

[17] ———, *Some enumerative results in the theory of forms*, Proc. Cambridge Philos. Soc., 39 (1943), pp. 22–30.

[18] J. HUMPHREYS, *Introduction to Lie Algebras and Representation Theory*, Springer-Verlag, New York, 1972.

[19] G. JAMES, *A characteristic-free approach to the representation theory of S_n*, J. Algebra, 46 (1977), pp. 430–450.

[20] R. KING, *Weight multiplicities for the classical groups*, in Group Theoretical Methods in Physics, Lecture Notes in Physics 50, Springer-Verlag, New York, 1975, pp. 490–499.

[21] ———, *The character generator of $Sp(2k)$*, preprint, Univ. of Southampton, 1981.

[22] V. LAKSHMIBAI, C. MUSILI AND C. SESHADRI, *Geometry of G/P-IV (Standard monomial theory for classical types)*, Proc. Indian Acad. Sci., 88A (1979), pp. 279–362.

[23] G. LANCASTER AND J. TOWBER, *Representation-functors and flag-algebras for the classical groups, I, II*, J. Algebra, 59 (1979), 16–38; J. Algebra, in press.

[24] F. MACAULAY, *Some properties of enumeration in the theory of modular systems*, Proc. London Math. Soc., 26 (1927), pp. 531–555.

[25] R. PROCTOR, *Solution of two difficult combinatorial problems with linear algebra*, preprint, Massachusetts Institute of Technology, Cambridge, MA, 1979.

[26] F. SCHWEINS, *Theorie der Differenzen und Differentiale, usw.*, vol. 3: *Theorie der Producte mit Versetzungen*, Heidelberg, 1825, pp. 345–355; as discussed in Muir, Theory of Determinants in the Historical Order of Development, second ed., Macmillan, London, 1906, vol. 1, chapter 6, pp. 159–175.

[27] R. STANLEY, *Weyl groups, the hard Lefschetz theorem, and the Sperner property*, this Journal, 1 (1980), pp. 180–184.

[28] J. SYLVESTER, *On homogeneous quadratic polynomials*, Philos. Mag., 4, No. II (1851), pp. 142–145.

[29] J. TOWBER, *Two new functors from modules to algebras*, J. Algebra, 47 (1977), pp. 80–104.

[30] ———, *Young symmetry, the flag manifold and representations of $GL(n)$*, J. Algebra, 61 (1979), pp. 414–462.

[31] H. TURNBULL, *The Theory of Determinants, Matrices and Invariants*, Blackie, London/Glasgow, 1929.

[32] K.-P. VO, *The Schensted correspondence and lexicographic matchings in multi-subset lattices*, this Journal, 2 (1981), pp. 324–332.

[33] H. WEYL, *Lecture notes*, Institute for Advanced Study, Princeton, NJ, 1934 (unpublished).

[34] ———, *The Classical Groups*, Princeton Univ. Press, Princeton, NJ, 1939.

[35] A. YOUNG, *On quantitative substitutional analysis* (second paper), Proc. London Math. Soc., (1) 34 (1902), pp. 261–397; (third paper), Proc. London Math. Soc., (2) 28 (1927), pp. 255–292.