SIAM J. ALG. DISC. METH.
Vol. 2, No. 1, March 1981

0196-5212/81/0201-0001 $01.00/0

ON THE DECOMPOSITION OF GRAPHS*

F. R. K. CHUNG†

Abstract. In this paper, we study the decompositions of a graph G into edge-disjoint subgraphs all of which belong to a specified class of graphs $\mathcal{H}$. Let $\alpha(G; \mathcal{H})$ denote the minimum value of the total sum of the sizes of subgraphs in $\mathcal{H}$ into which G can be decomposed, taken over all such decompositions of G. Let $\alpha(n; \mathcal{H})$ denote the maximum value of $\alpha(G; \mathcal{H})$ over all graphs G with n vertices.

In this paper, we settle a conjecture of Katona and Tarján by showing

$$\alpha(n; \mathcal{K}) = \lfloor n^2/2 \rfloor,$$

where $\mathcal{K}$ denotes the set of all complete graphs. Moreover, we show that the complete bipartite graph G on $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$ vertices is the only graph with $\alpha(G; \mathcal{K}) = \alpha(n; \mathcal{K})$.

I. Introduction. Many interesting problems in graph theory[1] can be described in the following general framework.

Suppose G is a finite connected graph with vertex set $V(G)$ and edge set $E(G)$. Consider a decomposition of G into subgraphs $G_1, G_2, \cdots, G_t$, such that any edge in G is an edge of exactly one of the G_i's, and all G_i's belong to a specified class of graphs $\mathcal{H}$. Such a decomposition will be called an *$\mathcal{H}$-decomposition* of G.

Let f be a *cost function* of graphs which assigns certain nonnegative values to all graphs. It is often of interest to consider the $\mathcal{H}$-decomposition of a given graph so that the total "cost" (i.e., the sum of the f values of all subgraphs in the $\mathcal{H}$-decomposition) is minimized. We define

$$\alpha_f(G; \mathcal{H}) = \min_P \sum_i f(G_i),$$

where $P = \{G_1, \cdots, G_t\}$ ranges over all $\mathcal{H}$-decompositions of G.

Typical questions one asks are to find $\alpha_f(G; \mathcal{H})$ or to determine

$$\alpha_f(n; \mathcal{H}) = \min_G \alpha_f(G; \mathcal{H}),$$

where G ranges over all graphs on n vertices.

Before proceeding to our main results, we shall first survey some of the known related results in this area. We abbreviate $\alpha(G; \mathcal{H}) = \alpha_{f_1}(G; \mathcal{H})$ and $\alpha(n; \mathcal{H}) = \alpha_{f_1}(n; \mathcal{H})$ where $f_1(G) = |V(G)|$. We also write $\alpha_*(G; \mathcal{H}) = \alpha_{f_0}(G; \mathcal{H})$ and $\alpha_*(n; \mathcal{H}) = \alpha_{f_0}(n; \mathcal{H})$ where $f_0(G) = 1$ for any graph G.

Let $\mathcal{B}$ denote the set of all complete biparite graphs. The problem of determining $\alpha(n; \mathcal{B})$ arises in the study of the networks of contacts realizing certain symmetric monotone Boolean functions (see [9], [13] and [16]). For this problem G. Hensel [9] obtained the estimate

$$n \log_2 n \leqq \alpha(K_n; \mathcal{B}) \leqq n \log_2 n + (1 - \log_2 e + \log_2 \log_2 e)n,$$

where e is the base of the natural logarithm. This question was also investigated by P. Erdös, A. Rényi and V. T. Sós [5], and the first part of the preceding inequality was also proved independently by G. Katona and E. Szemerédi [11].

* Received by the editors April, 1980, and in final form June 3, 1980.
† Bell Laboratories, Murray Hill, New Jersey 07974.
[1] The reader is referred to [8] for undefined terminology.

The bounds for $\alpha(n; \mathcal{B})$ were found by F. R. K. Chung, P Erdös and J. Spencer [2]; namely,

$$(1-\varepsilon)\frac{n^2}{2e \log n} < \alpha(n; \mathcal{B}) < (1+\varepsilon)\frac{n^2}{\log n},$$

for a given ε and large n, where $e = 2.718\cdots$.

A theorem of R. L. Graham and H. O. Pollak [6] asserts that for the complete graph K_n on n vertices,

$$\alpha_*(K_n; \mathcal{B}) = n - 1.$$

It is easily seen that a graph on n vertices can be decomposed into $n-1$ stars. Thus, we have

$$\alpha_*(n; \mathcal{B}) = n - 1.$$

Let $\mathcal{B}^*$ be the set of all bipartite graphs. It can be easily verified that

$$\alpha(n; \mathcal{B}^*) = \alpha(K_n; \mathcal{B}^*) = \alpha(K_n; \mathcal{B}) = \alpha(n; \mathcal{B}),$$

and

$$\alpha_*(n; \mathcal{B}^*) = \alpha_*(K_n; \mathcal{B}^*) = \lceil \log_2 n \rceil,$$

where $\lceil x \rceil$ denotes the least integer greater than or equal to x.

Let $\mathcal{F}$ denote the class of all forests, (i.e., acyclic graphs). In this case $\alpha_*(G, \mathcal{F})$ is usually called the *arboricity* of G (see [8]). Nash-Williams [17] gives the following expression for $\alpha(G; \mathcal{F})$:

$$\alpha_*(G; \mathcal{F}) = \max_S \left\lceil \frac{|E(S)|}{|V(S)|-1} \right\rceil,$$

when S ranges over all nontrivial induced subgraphs of G.

It is immediate that

$$\alpha_*(n; \mathcal{F}) = \alpha_*(K_n; \mathcal{F}) = \lceil n/2 \rceil.$$

Let $\mathcal{T}$ denote the class of all trees. F. R. K. Chung [1] showed that

$$\alpha_*(n; \mathcal{T}) = \lceil n/2 \rceil.$$

It can be easily seen that

$$\alpha(G; \mathcal{T}) = |E(G)| + \alpha_*(G; \mathcal{T})$$

and

$$\alpha(G; \mathcal{T}) \geqq \alpha(G; \mathcal{F}).$$

Thus we have

$$\alpha(n; \mathcal{T}) = \alpha(n; \mathcal{F}) = \lceil n^2/2 \rceil.$$

Finally, we should mention the striking work of R. M. Wilson [18] who investigated the decomposition of the complete graph into subgraphs which are all isomorphic to a specified graph, i.e., the case $G = K_n$ and $\mathcal{H} = \{H\}$. If such an $\mathcal{H}$-decomposition of G exists, then it follows immediately that: (a) the number of edges in H divides the number of edges in K_n;(b) the greatest common divisor of the degrees of vertices in H

divides $n-1$. Wilson showed that these two necessary conditions are sufficient for n sufficiently large (as a function of H).

A 15 year-old conjecture of T. Gallai asserts that for $\mathscr{P}$, the set of all paths, the following equality holds:

CONJECTURE (Gallai). $\alpha_*(n;\mathscr{P}) = \lceil n/2 \rceil$.

Let $\mathscr{C}$ denote the set of all simple cycles. G. Hajós conjectured that any graph on n vertices having all degrees even can always be decomposed into $\lfloor n/2 \rfloor$ or fewer simple cycles. For a graph G containing vertices having odd degree, we set $\alpha_*(G;\mathscr{C}) = 0$. We can write Hajós's conjecture as follows:

CONJECTURE (Hajós). $\alpha_*(n;\mathscr{C}) = \lfloor n/2 \rfloor$, *where* $\lfloor x \rfloor$ *denotes the greatest integer less than or equal to* x.

L. Lovász [15] proved a variation of the above conjecture by showing

$$\alpha_*(n;\mathscr{H}) = \lfloor n/2 \rfloor,$$

where $\mathscr{H}$ is the class of all paths and cycles. P. Erdös, A. W. Goodman and L. Pósa showed in [4] that a graph on n vertices can always be decomposed into $\lfloor n^2/4 \rfloor$ complete subgraphs, i.e., for $\mathscr{K}$, the set of all complete subgraphs,

$$\alpha_*(n;\mathscr{K}) = \lfloor n^2/4 \rfloor.$$

In fact, they sharpened the above result by showing

$$\alpha_*(n;\{K_2, K_3\}) = \lfloor n^2/4 \rfloor.$$

Finally G. Katona and T. Tarján [12] conjectured that

$$\alpha(n;\mathscr{K}) = \lfloor n^2/2 \rfloor.$$

In this paper, we prove this conjecture. Moreover, we show that the complete bipartite graph on $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$ vertices, denoted by B_n, is the only graph with $\alpha(G;\mathscr{K}) = \alpha(n;\mathscr{K})$.

2. On $\alpha(n;\mathscr{K})$ and $\alpha_*(n;\mathscr{K})$. First we remark that $\alpha(G;\mathscr{K})$ can still be defined (in the obvious way) if G is not connected. In the remaining part of the paper, the graphs we consider are not necessarily connected.

The main theorem of the paper will be the following:

THEOREM. *Any graph on n vertices can be decomposed into complete subgraphs so that the sum of the sizes of all subgraphs in this decomposition does not exceed* $\lfloor n^2/2 \rfloor$. *That is,*

(1)
$$\alpha(n;\mathscr{K}) = \max_{|V(G)|=n} \alpha(G;\mathscr{K}) = \lfloor n^2/2 \rfloor.$$

The only graph on n vertices satisfying $\alpha(G;\mathscr{K}) = \lfloor n^2/2 \rfloor$ *is the complete bipartite graph* B_n.

Proof. It is easy to see that the complete bipartite graph B_n has the property that

$$\alpha(B_n;\mathscr{K}) = \lfloor n^2/2 \rfloor.$$

To prove $\alpha(n;\mathscr{K}) = \lfloor n^2/2 \rfloor$, it suffices to show that for any graph on n vertices, we have

(2)
$$\alpha(G;\mathscr{K}) \leqq \lfloor n^2/2 \rfloor.$$

Let $G_1, \cdots, G_t$ be a decomposition of G into complete subgraphs. Let p_j denote the number of G_i's which are isomorphic to the complete graph on j vertices (denoted by K_j).

Thus,

$$\alpha(G;\mathcal{K}) \leqq \sum_{i=2}^{n} ip_i = 2e - \sum_{i=3}^{n} p_i i(i-2),$$

since

$$\sum_{i=2}^{n} p_i \binom{i}{2} = |E(G)| = e.$$

We note that inequality (2) holds if and only if there exists a $\mathcal{K}$-decomposition of G which satisfies the following:

$$2e - \sum_{i=3}^{n} p_i i(i-2) \leqq \lfloor n^2/2 \rfloor. \tag{3}$$

It is easily seen that the theorem holds for $n = 1$ or 2. We may assume $n \geqq 3$, and for any graph H on m vertices, $m < n$, we have

$$\alpha(H;\mathcal{K}) \leqq \lfloor m^2/2 \rfloor,$$

with equality if and only if H is B_m.

Let v^* be a vertex of G such that the degree of v^* does not exceed the degree of any other vertex in G; i.e.,

$$\deg v^* = \delta = \min_{v \in V(G)} \deg v.$$

Let L denote the induced subgraph on the set of vertices of G which are adjacent to v^*. A *vertex decomposition* of L is defined to be a set of vertex-disjoint complete subgraphs of L, say $M_1, M_2, \cdots, M_r$, such that $\sum_i |V(M_i)| = |V(L)|$. A vertex decomposition $M_1, \cdots, M_r$, where $|V(M_1)| \geqq |V(M_2)| \geqq \cdots \geqq |V(M_r)|$, of L is said to be *maximal* if for any vertex decomposition $N_1, \cdots, N_t$ either we have $|V(M_i)| = |V(N_i)|$ for $i = 1, \cdots, r$ where $r = t$, or there exists k such that $|V(M_k)| > |V(N_k)|$ and $|V(M_j)| = |V(N_j)|$ for $j < k$. Lst X_i denote the set of all complete subgraphs on i vertices in a fixed maximal vertex decomposition $P = \{M_1, \cdots, M_r\}$ of L.

We consider the graph G' with vertex set $V(G) - \{v^*\}$ and edge set $\{\{u, v\} \in E(G)$: $v^* \notin \{u, v\}$, and $\{u, v\}$ is not an edge of any M_i, $i = 1, \cdots, r\}$.

By the induction assumptions and (3), there exists a decomposition of G' into complete subgraphs, p_i' of which are isomorphic to K_i, $i = 2, \cdots, n-1$, such that

$$\sum_{i=3}^{n-1} p_i' i(i-2) \geqq 2e' - \lfloor (n-1)^2/2 \rfloor, \tag{4}$$

where

$$e' = |E(G')| = e - \delta - \sum_{i=2}^{\delta} x_i \binom{i}{2}$$

and x_i denotes $|X_i|$.

We consider a decomposition P^* of G consisting of the union of the preceding decomposition of G' and x_i complete subgraphs isomorphic to K_{i+1} with v^* as one of the vertices, $1 \leqq i \leqq n$.

The number of subgraphs in P^* which are isomorphic to K_i is just

$$p_i = p_i' + x_{i-1} \quad \text{for } 2 \leqq i \leqq n.$$

We want to show that this choice of p_i satisfies (3). We have

$$\begin{aligned}\sum_{i=3}^{n} p_i\, i(i-2) &= \sum_{i=3}^{n} (p_i' + x_{i-1}) i(i-2) \\ &\geqq 2e' - \lfloor (n-1)^2/2 \rfloor + \sum_{i\geqq 1} x_i (i+1)(i-1) \\ &= 2\left(e - \delta - \sum_{i\geqq 2} x_i \binom{i}{2}\right) - \lfloor (n-1)^2/2 \rfloor + \sum_{i\geqq 1} x_i (i^2-1) \\ &= 2e - \lfloor n^2/2 \rfloor + \left(\lfloor n^2/2 \rfloor - \lfloor (n-1)^2/2 \rfloor - 2\delta + \sum_{i\geqq 1} (i-1) x_i \right).\end{aligned} \tag{5}$$

In order to establish (3) it suffices to show that

$$\begin{aligned}\sum_{i\geqq 1} (i-1) x_i &\geqq 2\delta - \lfloor n^2/2 \rfloor + \lfloor (n-1)^2/2 \rfloor \\ &= 2\delta - n + \varepsilon_n,\end{aligned} \tag{6}$$

where

$$\varepsilon_n = \begin{cases} 0 & \text{if } n \text{ is even,} \\ 1 & \text{if } n \text{ is odd.} \end{cases}$$

We note that (6) is obviously true if $2\delta - n + \varepsilon_n < 0$. We may assume that $2\delta - n + \varepsilon_n \geqq 0$.

We consider the subgraph L. By the minimality of δ it is easy to see that any vertex in L has degree at least $2\delta - n$ in L. Let u^* be an arbitrary vertex of M_r. By the maximality of P, u^* is adjacent to at most $j-1$ vertices of any subgraph in X_j. Therefore, we have

$$2\delta - n \leqq \deg_L u^* \leqq \sum_{i\geqq 1} (i-1) x_i. \tag{7}$$

Consider first the case where n is even. Then (6) follows from (7) so that (3) is established and (2) is valid. Now, suppose G is a graph with $\alpha(G; \mathcal{H}) = \lfloor n^2/2 \rfloor$. We consider the $\mathcal{H}$-decomposition P^* of G. It is easily seen that

$$2e = \sum_{i\geqq 3} p_i i(i-2) = \sum_{i\geqq 2} i p_i \geqq \alpha(G; \mathcal{H}) = \lfloor n^2/2 \rfloor.$$

Thus, equality in (4), (5) and (6) holds. By the induction assumptions, G' is isomorphic to B_{n-1}. Hence δ is at most $\lceil (n-1)/2 \rceil$. From the equality in (6), δ is at least $n/2$. Therefore we have

$$\delta = n/2 \quad \text{and} \quad \sum_{i\geqq 1} (i-1) x_i = 0.$$

Thus $x_i = 0$ for all $i > 1$ and L is the trivial graph on $n/2$ vertices. Therefore G is B_n.

Next, consider the case where n is odd. From (6) and (7), we note that (3) holds unless

$$2\delta - n = \deg_L u^* = \sum_{i\geqq 1} (i-1) x_i,$$

and therefore we will assume the equality in (7).

Suppose equality in (4) holds. By the induction assumption, G' is, in fact, isomorphic to B_{n-1}. Since δ is the minimum degree in G, δ is at most $(n-1)/2$. However, we are assuming that $2\delta - n + \varepsilon_n \geqq 0$. Thus $2\delta - n + 1 = 0$ and inequality (3) holds. We may assume equality in (4) does not hold. This implies that the equality in (5) also becomes strict. Therefore we have

$$\begin{aligned} \sum_{i=3}^{n} p_i i(i-2) &\geqq 2e - \lfloor n^2/2 \rfloor + 1 + \sum_{i \geqq 1} (i-1)x_i - 2\delta + n - 1 \\ &= 2e - \lfloor n^2/2 \rfloor. \end{aligned} \tag{8}$$

Thus (3) is valid for n odd.

Suppose G is a graph satisfying $\alpha(G; \mathcal{K}) = \lfloor n^2/2 \rfloor$ and n is odd. We consider the following two possibilities:

(a) $\alpha(G'; \mathcal{K}) = \lfloor (n-1)^2/2 \rfloor$. It follows that equality in (4) holds. Thus G' is isomorphic to B_{n-1} and δ is equal to $(n-1)/2$. Equality in (5) and (6) also holds. Therefore

$$\sum_{i \geqq 1} (i-1)x_i = 0.$$

Thus, $x_i = 0$ for $i > 1$ and G is isomorphic to B_n.

(b) $\alpha(G'; \mathcal{K}) < \lfloor (n-1)^2/2 \rfloor$. Therefore equality in (4) does not hold. However, it follows from $\alpha(G; \mathcal{K}) = \lfloor n^2/2 \rfloor$ that equality in (8) holds. Thus

$$\sum_{i \geqq 1} (i-1)x_i = 2\delta - n = \deg_L u^*, \tag{9}$$

$$\alpha(G'; \mathcal{K}) = \lfloor (n-1)^2/2 \rfloor - 1. \tag{10}$$

In Case (b) we will prove a sequence of claims to establish that G is a complete t-partite graph, $t \geqq 3$. This will then show by Lemmas 1 to 4 that G does not satisfy $\alpha(G; \mathcal{K}) = \lfloor n^2/2 \rfloor$.

Since we assume the equality in (7), it follows immediately that any vertex in M_r is then adjacent to exactly $j-1$ vertices of any graph in X_j. Moreover, based on the fact that $\sum_{i \geqq 1} ix_i = \delta$, we have $\sum_{i \geqq 1} x_i = n - \delta = r$.

CLAIM 1. *Any vertex in L is adjacent to at least $|V(M_r)| - 1$ vertices of M_r.*

Proof. Suppose to the contrary that a vertex w in L is not adjacent to u and u' in M_r. Assume w is a vertex of M_i for some i. Thus u and u' are adjacent to all vertices of M_i except for w. Then the induced graph of L on $(V(M_i) - \{w\}) \cup \{u, u'\}$ is a complete graph with more vertices than M_i. This contradicts the maximality of P. □

CLAIM 2. *Let u_i be a vertex of M_i such that u_i is not adjacent to $u^* = u_r$ for $i = 1, \cdots, r-1$. Then u_i, $1 \leqq i \leqq r$, is adjacent to exactly $j-1$ vertices of any graph in X_j.*

Proof. Suppose u_i is adjacent to all vertices in M_t. It follows that $r > t > i$. We consider another vertex decomposition of L, called $P' = \{N_1, \cdots, N_r\}$, such that $N_j = M_j$ if $j \neq i, t, r$; N_i is the induced subgraph of L on $(V(M_i) - \{u_i\}) \cup \{u^*\}$; N_t is the induced subgraph of L on $V(M_t) \cup \{u_i\}$; and N_r is the induced subgraph of L on $V(M_r) - \{u^*\}$. Thus P is not maximal. This is impossible. Therefore u_i, $1 \leqq i \leqq r$ is adjacent to at most $j-1$ vertices of any graph in X_j. Since

$$\deg_L u_i \geqq 2\delta - n = \sum_{j \geqq 1} (j-1)x_j,$$

we conclude that u_i is adjacent to exactly $j-1$ vertices of any graph in X_j. □

CLAIM 3. *u_i, $i = 1, \cdots, r$, is adjacent to every vertex in $V(G) - V(L)$ and the degree of u_i in G is δ.*

Proof. It follows from Claim 2 that the degree of u_i is at most

$$r + \sum_{j \geq 1} (j-1)x_j = (n-\delta) + (2\delta - n) = \delta.$$

On the other hand, δ is the minimum degree in G. Thus the degree of u_i in G is δ and u_i is adjacent to any vertex in $V(G) - V(L)$. □

CLAIM 4. *Let w be a vertex in L which is not adjacent to u_i for some i. Then w is of degree δ.*

Proof. Suppose $i = r$. It follows from Claim 3 that w has degree δ. We may assume $i \neq r$. We can also assume $w \neq u_i$ and w is not a vertex of M_i. Let w be a vertex of M_j. Suppose $j = r$. Then w has degree δ. We consider the case $j \neq r$. Suppose w is adjacent to all vertices in M_t. If $t \neq r$, we consider the following vertex decomposition $P'' = \{L_1, \cdots, L_r\}$ such that $L_k = M_k$ if $k \neq i, j, t, r$; L_i is the induced subgraph of L on $(V(M_i) - \{u_i\}) \cup \{u^*\}$; L_j is the induced subgraph of L on $(V(M_j) - \{w\}) \cup \{u_i\}$; L_t is the induced subgraph of L on $V(M_t) \cup \{w\}$; and L_r is the induced subgraph of L on $V(M_r) - \{u^*\}$. This contradicts the maximality of P. Thus, w is adjacent to at most $j-1$ vertices of any graph in X_j. If $t = r$, we consider the following vertex decomposition $\bar{P}'' = \{L_1', \cdots, L_r'\}$ such that $L_k' = M_k$ if $k \neq i, j, r$; L_i' is the induced subgraph of L on $(V(M_i) - \{u_i\}) \cup \{u^*, w\}$; L_j' is the induced subgraph of L on $(V(M_j) - \{u_j\}) \cup \{u_i\}$; and L_r' is the induced subgraph of L on $V(M_r) - \{u^*\}$. This again contradicts the maximality of P. Since the degree of w is at least δ in G and $2\delta - n$ in L, we conclude that the degree of w is δ in G and w is adjacent to exactly $j-1$ vertices of any graph in X_j. □

CLAIM 5. *$|V(M_i)| = |V(M_1)|$ for $i = 1, \cdots, r$.*

Proof. We choose w in G' with minimum degree δ' in G'. Let L' be the induced subgraph on the set of vertices of G' which are adjacent to w. Let $M_1', M_2', \cdots, M_s'$ be a maximal vertex decomposition of L'. We consider G'' with vertex set $V(G') - \{w\}$ and edge set $\{(u, v) \in E(G'): w \notin \{u, v\}$ and (u, v) is not an edge of any M_i', $i = 1, \cdots, s\}$.

By the induction assumptions, there exists a decomposition of G'' into complete subgraphs, p_i'' of which are isomorphic to K_i, $i = 2, \cdots, n-2$, such that

$$\sum_{i=3}^{n-2} p_i'' i(i-2) = 2e'' - \alpha(G''; \mathcal{K}) \geq 2e'' - \lfloor (n-2)^2/2 \rfloor, \tag{11}$$

where

$$e'' = |E(G'')| = e' - \delta' - \sum_{i=2}^{\delta'} x_i' \binom{i}{2},$$

and x_i' denotes the cardinality of the set X_i' which consists of all complete subgraphs on i vertices in the maximal vertex decomposition $P' = \{M_1', \cdots, M_s'\}$ of L'.

We consider a decomposition of G' consisting of the union of P'' and x_i' complete subgraphs isomorphic to K_{i+1} with w as one of its vertices, $1 \leq i \leq \delta'$. Let q_i denote the number of subgraphs in this decomposition of G' which are isomorphic to K_i. From (10) we have

$$2e' - \lfloor (n-1)^2/2 \rfloor + 1 \geq \sum_{i=3}^{n-2} q_i i(i-2).$$

From (11) and $\sum_{i\geqq 1} ix'_i = \delta'$, we also have

$$\sum_{i=3}^{n-2} q_i i(i-2) = \sum_{i=3}^{n-2} (p''_i + x'_{i-1}) i(i-2)$$

$$= 2e' - \alpha(G''; \mathcal{H}) - \delta' - \sum_{i\geqq 1} x'_i.$$

By the maximality of P', we have

$$2\delta' - (n-1) \leqq \text{minimum degree in } L' \leqq \sum_{i\geqq 1} (i-1)x'_i. \tag{12}$$

Therefore we have

$$n-1-\delta' \geqq \sum_{i\geqq 1} x'_i \geqq \lfloor (n-1)^2/2 \rfloor - \alpha(G''; \mathcal{H}) - \delta' - 1, \tag{13}$$

i.e., $\alpha(G''; \mathcal{H}) \geqq \lfloor (n-2)^2/2 \rfloor - 1$.

Suppose $\alpha(G''; \mathcal{H}) = \lfloor (n-2)^2/2 \rfloor$. Then by the induction assumptions, G'' is B_{n-2} and δ' is at most $(n-1)/2$. On the other hand, δ' is at least $(n-1)/2$. (Suppose $\delta' \leqq (n-1)/2 - 1$. Since $\sum_{i\geqq 1} ix'_i = \delta'$, we have $n-3-\delta' \geqq \sum_{i\geqq 1} x'_i$. From (13), we will then have $\alpha(G''; \mathcal{H}) > \lfloor (n-2)^2/2 \rfloor$, which contradicts the induction assumptions.) Therefore $\delta' = (n-1)/2$. From (13), we also have $\sum_{i\geqq 1} x'_i \geqq n-\delta'-2$, i.e., $\sum_{i\geqq 1} (i-1)x'_i \leqq 1$. Suppose $\sum_{i\geqq 1} (i-1)x'_i = 0$. Then G' is B_{n-1}, which contradicts (10). Therefore we have $x'_2 = 1$ and $x'_1 = \delta' - 2$. It can be easily verified that G' has a $\mathcal{H}$-decomposition which contains one K_3 and $(n-1)^2/4 - 3$ K_2. This contradicts (10). Thus we may assume $\alpha(G''; \mathcal{H}) = \lfloor (n-2)^2/2 \rfloor - 1$, and

$$\sum_{i\geqq 1} (i-1)x'_i = 2\delta' - n + 1 = \text{minimum degree in } L'. \tag{14}$$

We note that Claim 1 to Claim 4 all follow from (9) and the maximality of P. Therefore we can show in a similar manner that for any vertex w with degree δ' there exists a vertex w^* with degree δ' in G', so that the $n-1-\delta'$ vertices which are not adjacent to w^* are of degree δ' in G' and w adjacent to w^* in G'.

Let k be the size of $V(M_1)$. Then it can be easily seen that $\delta' = \delta - k$. We also note that any vertex in M_i has degree at least $\delta - |V(M_i)|$ in G'.

Since u_1 has degree δ' in G', there is a vertex $\bar{w}$ which is adjacent to u_1 in G' such that all the $n-1-\delta'$ vertices which are not adjacent to $\bar{w}$ are of degree δ' in G. Since any vertex in $G'-L$ had degree at least δ which is greater than δ', we may assume $\bar{w}$ is a vertex of $\bar{M} = M_i$, where $|V(M_i)| = k$. Since all vertices of $\bar{M}$ are not adjacent to $\bar{w}$ in G', all vertices of $\bar{M}$ have degree δ' in G'. Without loss of generality, we may assume that all vertices in M_1 have degrees δ' in G'.

Now, let M^* be the set of vertices in M_1 which are not adjacent to some vertex in M_r. Since every vertex in M_r is adjacent to exactly $k-1$ vertices in M_1 and any vertex in M_1 is adjacent to at least $k'-1$ vertices in M_r, where $|V(M_r)| = k'$, we have $|M^*| = |V(M_r)|$. Suppose M^* is a proper subset of $V(M_1)$. We may choose w to be a vertex in $V(M_1) - M^*$. Thus u^* is in L'. From Claims 1 to 4, there exists a vertex w^* with degree δ' in G' so that the $n-1-\delta'$ vertices which are not adjacent to w^* are of degree δ' in G'. We may assume, without loss of generality, that w^* is a vertex of M_2. Either u^* is nonadjacent to w^* or u^* is nonadjacent to a vertex in M_2 which is nonadjacent to w^* in L'. Thus, by Claims 3 and 4, the degree of u^* is δ' in G'. This implies $|V(M_r)| = k = |M^*|$.

This contradicts our assumption that M^* is a proper subset of $V(M_1)$. Therefore, we have shown that $|V(M_i)| = |V(M_1)| = k$.

CLAIM 6. *Each vertex in L is adjacent to exactly $j-1$ vertices of a graph in X_j and has degree δ.*

Proof. This follows from Claims 2 and 5. □

CLAIM 7. *For any i, j, $1 \leqq i, j \leqq r$, u_i and u_j are not adjacent to each other.*

Proof. Suppose u_i and u_j are adjacent. u_j is adjacent to exactly $k-1$ vertices of M_i. Let w denote the vertex in M_i which is not adjacent to u_j. From Claim 6, we have that w is adjacent to all vertices in M_j except for u_j. Now, we consider a vertex decomposition $\bar{P} = \{R_1, \cdots, R_r\}$ of L, where $R_t = M_t$ if $t \neq i, j, r$; M_j is the induced subgraph of L on $V(M_j) - \{u_j\} \cup \{u^*, w\}$; M_i is the induced subgraph of L on $V(M_i) - \{u_i\}$; M_r is the induced subgraph of L on $V(M_r) - \{u^*\} \cup \{u_i\}$. This contradicts the maximality of P. □

CLAIM 8. *Any vertex v in L has the property that the $n-\delta-1$ vertices which are not adjacent to v are mutually nonadjacent.*

Proof. This follows from Claims 5 and 7. □

CLAIM 9. *Any vertex v in G has the property that the degree of v is δ and the $n-\delta-1$ vertices which are not adjacent to v are mutually nonadjacent.*

Proof. This follows from Claim 8 and the fact that the choice of v^* is arbitrary. □

CLAIM 10. *G is a complete t-partite graph, where $t = n/(n-\delta) \geqq 3$. $V(G)$ is a disjoint union of t sets of cardinality $n-\delta$, and two vertices in G are adjacent if and only if they do not belong to the same set.*

Proof. Let I_1 be the set of vertices of G each of which is v_1 or is not adjacent to v_1. It follows from Claim 9 that any vertex in I_1 is adjacent to all vertices not in I_1 and not adjacent to any vertex in I_1 in G. If I_1 is a proper subset of $V(G)$, we choose a vertex v_2 in $V(G) - I_1$. Let I_2 be the set of vertices of G which is v_2 or is not adjacent to v_2. After a finite number of steps, we have sets $I_1, \cdots, I_t$ and G is a t-partite graph and $|I_i| = n-\delta$. Since $n \geqq 3$, we have $\alpha(G; \mathcal{K}) = [n^2/2] > 0$. Therefore G is not the trivial graph; i.e., $t > 1$. Since n is odd, t is not 2. We have $t \geqq 3$. □

We have shown that G is a complete t-partite graph where $t \geqq 3$. In the following there are some auxiliary lemmas dealing with the value $\alpha(G; \mathcal{K})$ for complete t-partite graphs G.

LEMMA 1. *Let Q be a complete t-partite graph on $V(Q) = I_1 \cup \cdots \cup I_t$ and $|I_i| = t$ for $i, \cdots, t$. Suppose t is a prime number. Then $\alpha(Q; \mathcal{K}) = t^3$.*

Proof. Let v_{ij}, $1 \leqq j \leqq t$, be vertices in I_i. Let ${}_jQ_k$ denote the complete graph on $v_{i,z}$, $i = 1, \cdots, t$, where $z \equiv (i-1)(k-1)+j \pmod t$ and $1 \leqq z \leqq t$. We note that $\{{}_jQ_k : 1 \leqq j, k \leqq t\}$ is a $\mathcal{K}$-decomposition of Q. Thus

$$\alpha(Q; \mathcal{K}) \leqq t^3.$$

On the other hand, the maximal complete subgraph contained in Q is K_t. Let P be a $\mathcal{K}$-decomposition of Q. For any edge e in Q, define w as follows:

$$w(e) = \frac{2}{f(e)-1},$$

where $f(e)$ is the number of vertices of the graph in P which contains e.

It is easy to see that

$$w(e) \geqq \frac{2}{t-1}.$$

We note that

$$\alpha(Q;\mathcal{K}) = \min_P \sum_e w(e) \geqq |E(Q)| \cdot \frac{2}{t-1} = \binom{t}{2} t^2 \cdot \frac{2}{t-1} = t^3.$$

Therefore, $\alpha(Q;\mathcal{K}) = t^3$. □

LEMMA 2. *Let Q be a complete 3-partite graph on $I_1 \cup I_2 \cup I_3$ and $|I_i| = t$ for $i = 1, 2, 3$. Then $\alpha(Q;\mathcal{K}) = 3t^2$.*

Proof. Let v_{ij}, $1 \leqq j \leqq t$, be vertices in I_i. We consider a $\mathcal{K}$-decomposition of Q which consists of the following graphs: $Q_{j,k}$, $1 \leqq j, k \leqq t$, where $Q_{j,k}$ is the complete graph on $v_{1,j}, v_{2,s}, v_{3,k}$ where $s \equiv k+j \pmod t$ and $1 \leqq s \leqq t$. Thus $\alpha(Q;\mathcal{K}) \leqq 3t^2$. By a similar proof to that in Lemma 1 we have $\alpha(Q;\mathcal{K}) = 3t^2$. □

LEMMA 3. *Let Q be a complete t-partite graph on $I_1 \cup I_2 \cup \cdots \cup I_t$ and $|I_i| = 3$ for $i = 1, 2, \cdots, t$. Then $\alpha(Q;\mathcal{K}) \leqq 3t^2$.*

Proof. We consider the following two possibilities:

Case 1. $t \neq 0 \pmod 2$. We consider a $\mathcal{K}$-decomposition consisting of

(a) Q^j, $1 \leqq j \leqq 3$, where Q^j is the complete graph on $v_{i,j}$, $i = 1, \cdots, t$; and

(b) Q_{jk}, $1 \leqq j \leqq t$, $1 \leqq k \leqq t$, $j \neq k$, where Q_{jk} is the complete graph on $v_{j,1}, v_{s,2}, v_{k,3}$ and $s \equiv \sigma^{-1}(j+k) \pmod t$ and σ is the permutation $\sigma(i) = 2i$.

Case 2. $t \equiv 0 \pmod 2$. We consider a $\mathcal{K}$-decomposition consisting of

(a) Q^j, $1 \leqq j \leqq 3$; and

(b) Q'_{jk}, $1 \leqq j \leqq t$, $1 \leqq k \leqq t$, $j \neq k$, where Q'_{jk} is the complete graph on $v_{j,i}, v_{s,2}, v_{k,3}$ and $s \equiv \zeta^{-1} f(j+k)$ where

$$f(x) = \begin{cases} x & \text{if } x \leqq t, \\ x-t-1 & \text{if } x > t+1, \\ t & \text{if } x = t+1, \end{cases}$$

and ζ is the permutation $f(i) = f(2i)$.

It is easy to check that in both cases we have $\mathcal{K}$-decompositions and

$$\alpha(Q;\mathcal{K}) \leqq 3t^2. \qquad \square$$

LEMMA 4. *Let G be a complete t-partite graph on $V(G) = I_1 \cup \cdots \cup I_t$ and $|I_i| = s$ for $1 \leqq i \leqq t$. Let $q = \max(s, t)$. We have*

$$\alpha(G;\mathcal{K}) \leqq n(2q-2),$$

where

$$|V(G)| = n = st.$$

Proof. By Bertrand's postulate (see [10]), there exists a prime p between q and $2q-2$. We will show that $\alpha(G;\mathcal{K}) \leqq np$.

It is easy to see that G is a subgraph of the complete p-partite graph Q on $I'_1 \cup \cdots \cup I'_p$ where $|I'_i| = p$. From the proof of Lemma 1 there is a $\mathcal{K}$-decomposition P of Q which consists of subgraphs isomorphic to K_p. We consider a $\mathcal{K}$-decomposition P' of G which is defined to be $\{G_i \cap G : G_i \in P\}$. Therefore, we have

$$\alpha(G;\mathcal{K}) \leqq \sum_{G_i \in P} |V(G_i \cap G)| \leqq np \leqq n(2q-2),$$

since every vertex is in at most p of the K_p's. □

Now, from Lemmas 2 and 3 and the fact that $3t^2 < \lfloor (3t)^2/2 \rfloor$ for t positive, we may assume $r = n - \delta > 3$, $t = n/(n-\delta) > 3$. Since $rt = n$, we have $q \leqq n/4$. From Lemma 4, we have

$$\alpha(G; \mathcal{K}) \leqq n(2q-2) \leqq n(n/2-2).$$

This contradicts the assumption that $\alpha(G; \mathcal{K}) = \lfloor n^2/2 \rfloor$. Thus we have shown that Case (b) is impossible.

We also note that

$$2\alpha_*(G; \mathcal{K})) \leqq \alpha(G; \mathcal{K}).$$

Any graph G with $\alpha_*(G; \mathcal{K}) = \alpha_*(n, \mathcal{K}) = [n^2/4]$ must then have $\alpha(G; \mathcal{K}) = \lfloor n^2/2 \rfloor = \alpha(n; \mathcal{K})$. Therefore, a graph G on n vertices having $\alpha(G; \mathcal{K}) = \alpha(n; \mathcal{K})$ or $\alpha_*(G; \mathcal{K}) = \alpha_*(n; \mathcal{K})$ is the complete bipartite graph on $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$ vertices. This completes the proof of the main theorem. □

3. Concluding remarks. Problems of the type we have discussed in this paper are not only interesting in their own right, but also have potential applications in communication and switching networks. Sometimes it is desirable to decompose a communication or switching network into parts of certain specified types. The problem of determining $\alpha(G; \mathcal{H})$ or $\alpha_*(G; \mathcal{H})$ is equivalent to the problem of minimizing the "cost" of building the network (with the corresponding graph G) by using certain types of parts $\mathcal{H}$. In fact, the study of $\alpha(n, \mathcal{B})$ was first motivated by consideration of contact networks.

There are many interesting problems left open in this area. For example, the Gallai conjecture on $\alpha(n; \mathcal{P})$ still remains unsolved. We can ask the question of determining $\alpha(n; \mathcal{H})$ for $\mathcal{H}$ a class of graphs with certain specified properties, e.g., each graph has connectivity $\leqq \rho$, has chromatic number $\leqq \tau$, etc.

We can consider a variation of the above problems. For a graph G and a class of graphs $\mathcal{H}$, we define an $\mathcal{H}$-covering of G to be a family of subgraphs, $G_1', \cdots, G_{t'}'$, such that every edge is in at least one of the G_t' and every G_t' is in $\mathcal{H}$. For a cost function f, we define

$$\beta_f(G; \mathcal{H}) = \min_{P'} \sum_1 f(G_t'),$$

where $P' = \{G_1', \cdots, G_{t'}'\}$ ranges over all $\mathcal{H}$-coverings of G,
and

$$\beta_f(n; \mathcal{H}) = \max_G \beta_f(G; \mathcal{H}),$$

where G ranges over all graphs on n vertices.

We note that an $\mathcal{H}$-decomposition of G is also an $\mathcal{H}$-covering of G. Therefore

$$\beta_f(G; \mathcal{H}) \leqq \alpha_f(G; \mathcal{H}),$$

and

$$\beta_f(n; \mathcal{H}) \leqq \alpha_f(n; \mathcal{H}).$$

The preceding equalities sometimes hold and sometimes do not. For example, from the main result of this paper it is easily seen that

$$\alpha_{f_1}(n; \mathcal{K}) = \beta_{f_1}(n; \mathcal{K}) = \lfloor n^2/2 \rfloor.$$

However,

$$\beta_{f_0}(K_n;\beta)=\lfloor \log_2 n \rfloor \quad \text{and} \quad \alpha_{f_0}(K_n;\beta)=n-1.$$

Also from [3] we have

$$n-n^{11/14+\varepsilon}<\alpha_{f_0}(n;\beta)<n-c\log n,$$

for a given $\varepsilon>0$ and some constant c if n is sufficiently large.

We can ask the corresponding question of determining $\beta_f(G;\mathcal{H})$ or $\beta_f(n;\mathcal{H})$ for various classes of graph $\mathcal{H}$ and cost functions f.

We could also consider another kind of variation of this problem in which we wish to decompose a graph G into *induced* subgraphs of some certain type $\mathcal{H}$. We can then ask the corresponding question of determining $\alpha'_f(G;\mathcal{H})$, the minimum cost of subgraphs over all possible decompositions of G, and $\alpha'_f(n;\mathcal{H})$, the maximum value of $\alpha'_f(G;\mathcal{H})$ over all graphs G on n vertices.

Remarks. J. Kahn [7] proved that $\beta_{f_1}(n;\mathcal{H})=\lfloor n^2/2 \rfloor$. E. Győri and A. V. Kostočka [10] have recently proved the main result in this paper by a completely different method.

Acknowledgment. The author wishes to thank T. H. Foregger for his helpful comments.

REFERENCES

[1] F. R. K. CHUNG, *On Partitions of graphs into trees*, Discrete Math., to appear.
[2] F. R. K. CHUNG, P. ERDÖS AND J. SPENCER, *On the decomposition of graphs into complete bipartite subgraphs*, to appear.
[3] F. R. K. CHUNG, *On coverings of graphs*, to appear.
[4] P. ERDÖS, A. W. GOODMAN AND L. PÓSA, *The representation of graphs by set intersections*, Canad. J. Math., 18 (1966), pp. 106–112.
[5] P. ERDÖS, A. RÉNYI AND V. T. SÓS, *On a problem of graph theory*, Studia Sci. Math. Hungar., 1 (1966), pp. 215–235.
[6] R. L. GRAHAM AND H. O. POLLAK, *On the addressing problem for loop switching*, Bell System Tech. Jour., 50 (1971), pp. 2495–2519.
[7] J. KAHN, *Proof of a conjecture of Katona and Tarján*, to appear.
[8] F. HARARY, *Graph Theory*, Addison-Wesley, New York, 1969.
[9] G. HENSEL, *Nombre minimal de contacts de fermature nécessaires pour réaliser une fonction booléenne symétrique de n variables*, C. R. Acad. Sci. Paris, 258 (1964), pp. 6037–6040.
[10] E. GYÖRI AND A. V. KOSTOČKA, *On a problem of G. O. H. Katona and T. Tarján*, to appear.
[11] G. KATONA AND E. SZEMERÉDI, *On a problem of graph theory*, Studia Sci. Math. Hungar., 2 (1967), pp. 23–28.
[12] G. KATONA AND T. TARJAN, *personal communication.*
[13] R. E. KRICHEVSKII, *Complexity of contact circuits realizing a function of logical algebra*, Soviet Phys. Dokl., 8 (1964), pp. 770–772.
[14] E. LANDAU, *Handbuch der Lehre von der Verteilung der Primzahlen, I*, 1909, p. 89–92.
[15] L. LOVÁSZ, *On coverings of graphs*, in Theory of Graphs, P. Erdös and G. Katona, eds., Academic Press, New York, 1968, pp. 231–236.
[16] O. B. LUPANOV, *On comparing the complexity of the realizations of monotone functions by contact networks containing only closing contacts and by arbitrary contact networks*, Soviet Phys. Dokl., 7 (1962), pp. 486–489.
[17] C. ST. J. A. NASH-WILLIAMS, *Decomposition of finite graphs into forests*, J. London Math. Soc., 39 (1964), p. 12.
[18] R. M. WILSON, *Decompositions of complete graphs into subgraphs isomorphic to a given graph*, Proc. 5th British Combinatorial Conference 1975, pp. 647–659.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 1, March 1981

0196-5212/81/0201-0002 $01.00/0

A PROBLEM WITH TELEPHONES*

RICHARD T. BUMBY†

Abstract. This paper deals with the "telephone problem," also known as the "gossip problem". Suppose n persons each have a piece of information. Pairs of them can share whatever information they possess by making a telephone call. The question arises, what minimum number of calls allows all n persons to obtain all n pieces of information. The answer is $2n-4$. One can then ask about properties of such minimal sets of calls. In particular, we prove that the graph whose edges are the calls must contain a four-cycle.

1. Introduction. The "telephone problem" has often been solved in the literature [1], [3], [7], and various extensions of the problem have been proposed [2], [4], [5]. This paper is devoted to the proof of the "four-cycle conjecture" introduced in [4] with the words: "We are so convinced of the next statement that even though it is by definition a conjecture, we shall call it a *True conjecture*: $\cdots$" (italics theirs). The rumor of its solution, hastily added in proof in [4], proved to be premature.

We now establish the notation for the proof. We assume that there are n persons and k calls between them. The persons form a set U and the calls form an ordered set $T=\langle t_1, \cdots, t_k\rangle$ of (unordered) pairs of distinct elements of U. T is called a "system of calls."

It is natural to think of T as determining a graph $G(T)$ whose verticles are U and whose edges are the elements of T. Thus, the elements of U are also called "vertices" or "nodes."

The ordering on T can be used to introduce a relation $a \to b$ which holds iff there is a path $a=x_0, \cdots, x_m=b$ such that there is an increasing subsequence $\langle s_i: 1 \leqq i \leqq m\rangle$ in T and $s_i=\{x_{i-1}, x_i\}$. The relation $a \to b$ means that b learns a's information in the system of calls T. If $a \to b$, we then have a path between a and b in $G(T)$, so these points lie in the same component of $G(T)$.

DEFINITION. A system T is called *pooling* if $a \to b$ for each $a, b \in U$.

This paper proves:

THEOREM 1. *If T is pooling, then $k \geqq 2n-4$.*

THEOREM 2. *If T is pooling and $k=2n-4$, then $G(T)$ contains a four-cycle.*

2. The minimal partial ordering of T. We have described T as a *sequence* of calls; that is, the t_i are totally ordered in time. However, note that whether $a \to b$ holds depends only on the following partial order on T:

DEFINITION. The *minimal order* on T is the transitive closure of the relation $\{(t_i, t_j): i \leqq j \text{ and } t_i \cap t_j \neq \varnothing\}$.

Thus, two calls are ordered in the minimal order only if information can flow through one call into another, or equivalently, if it is the case that their order in time could not be reversed without changing information paths. If we consider all t_i containing a fixed node, the minimal order gives a linear order on them. Since all the essential properties of T are given by the minimal order, we will henceforth ignore the original linear ordering.

If we consider any linear ordering (time sequencing) of T that is compatible with the minimal ordering, and select a time between two of the calls, we partition T into the calls I before that time, and the calls F after that time. The pair (I, F) has the following nice properties:

* Received by the editors May 17, 1979, and in revised form June 13, 1980.

† Department of Mathematics, Hill Center for the Mathematical Sciences, Rutgers University, New Brunswick, New Jersey 08903.

(i) $T = I \cup F$,
(ii) $I \cap F = \varnothing$,
(iii) $i \in I$ and $t \leqq i$ imply $t \in I$,
(iv) $f \in F$ and $t \geqq f$ imply $t \in F$.

That is, I and F are complementary lower and upper ideals of the partially ordered set T. Conversely, if (I, F) has the properties (i)–(iv), then there is a time ordering of T that performs the calls I before the calls F.

DEFINITION. (I, F) satisfying (i)–(iv) above is a *splitting* of T.

If we give a splitting (I, F) and we have defined only one component, the other component is its complement in T.

3. Components and closures. If (I, F) is any splitting and u is a node that is involved in some call in F, we define $\min(u)$ to be the first call in F involving u. Clearly, if $u \in t \in F$, then $\min(u) \leqq t$. If $t = (u_1, u_2) \in F$, then $t \geqq \min(u_1), \min(u_2)$. Unless $t = \min(u_1) = \min(u_2)$, there is a call in F strictly smaller (earlier) than t. Also, if $t > t_0 \in F$, then information can flow from t_0 through an increasing sequence of calls to t. The last call of this chain before t must share a node u with t, and so $t \neq \min(u)$. This proves:

PROPOSITION 1. *For any splitting (I, F), a call $t = (u, v)$ of F is minimal in F iff $t = \min(u) = \min(v)$.*

DEFINITION. Given a splitting (I, F) with $I \geqq n-2$ and $G(I)$ not connected, any component of $G(I)$ is a *component of* T.

Starting from the splitting $(\varnothing, T)$, we can successively "bring down" minimal elements of F from F into I. Thus, we can construct splittings (I, F) in which I (or F) has any size from 0 to k. Clearly, any pooling must have $k \geqq n-1$, so selecting $|I| = n-2$ shows that all poolings have components.

The basis of our proofs will be constructions involving the components of T. The first construction will be closure.

Suppose that X is a component of T defined by the splitting (I_0, F_0). Define A_X to be the splitting (I, F) with the largest $I \subseteq I_0$ such that X is a component of $G(I)$. Define B_X to be the splitting (I, F) with the largest $I \subseteq I_0$ that has a component with the same vertices as X.

DEFINITION. If B_X is (I, F), then the component of $G(I)$ with the same vertices as X is the *closure* of X, denoted $\bar{X}$. If $\bar{X} = X$, $B_X = A_X$ and X is *closed*, and (I, F) is called the *canonical splitting* for X.

PROPOSITION 2. *Given a splitting (I, F), if no minimal element of F joins two nodes of X, then $A_X = B_X$, and hence, X is closed. In addition, if every minimal element of F joins a node of X to a node not in X, then $A_X = B_X = (I, F)$. Conversely, if (I, F) is the canonical splitting of a closed X, every minimal element of F joins a node of X to a node not in X.*

Proof. Let $A_X = (I_A, F_A)$ and $B_X = (I_B, F_B)$. I_B is I_A with some additional calls from F_A that connect nodes of X. But, if no minimal element of F joins two nodes of X, no such call can be moved into I_B without connecting X to some other component. If, in addition, every minimal element of F joins a node of X to a node not in X, then no call in F can be moved to I_A without connecting X to some other component. Conversely, if a minimal element of F connects two nodes of X, X is not closed, and if a minimal element of F connects two nodes not in X, then (I, F) is not the canonical splitting of X.

COROLLARY. *If X consists of a single point, then X is closed.*

We use the convention that all "lemmas" have the standing hypothesis that T is a pooling system on U. "Propositions" deal with general call systems.

LEMMA 1. *If any component X consists of a single point x, then $k \geqq 2n-3$. Hence, Theorems 1 and 2 hold in this case.*

Proof. From the corollary to Proposition 2, X is closed. Let (I, F) be its canonical splitting. Except in the trivial case $n = 1$, no node of U has received all the information yet, so each node u has a min (u). From Proposition 2, $\{\min(u): u \neq x\}$ are distinct, and so F has $\geqq n-1$ elements. Since $|I| \geqq n-2$, the result follows.

From Lemma 1, if $k = 2n-4$, then T can have no component consisting of a single point. Let $\min_0(u)$ be the first call involving u in T; i.e., $\min_0(u)$ is $\min(u)$ for the splitting $(\varnothing, T)$. Then for any splitting (I, F) with $|I| \geqq n-2$, $\min_0(u) \in I$ for any node u.

4. Minimal trees. Suppose we have a closed component X which is a tree. After Lemma 1, we may assume that $|X| > 1$. Let (I, F) be its canonical splitting. As X is a component of $G(I)$, I is a disjoint union $I_X \cup I_Y$, where I_X is all calls between nodes of X and I_Y is the rest. Let $t_0 = (x_0, x_1)$ be a maximal element of I_X. In $G(I - t_0)$, X falls into two parts X_0, X_1 with $x_i \in X_i (i = 0, 1)$.

Suppose F has a minimal element $t_1 = (x_2, y)$ with $y \notin X$ and $x_2 \neq x_0, x_1$. Then t_0 and t_1 are incomparable in T. If we bring t_1 down into I and raise t_0 up into F, we get a new splitting (I', F') with $|I'| = |I|$. What does $G(I')$ look like?

Say $x_2 \in X_0$; then the addition of t_1 to $I - t_0$ causes X_0 to be connected to the component containing y. As y must be outside X_1, this leaves X_1 as a component of $G(I')$. The minimal elements of F' are t_0, some minimal elements of F, and some elements of F having nodes in common with t_1. None of these can have both nodes in X_1. From Proposition 2, X_1 is closed, with canonical splitting (I'', F''), $I' \subseteq I''$.

Induction on this construction proves:

PROPOSITION 3. *If X is a closed tree component of T corresponding to a splitting (I, F), then we can find a closed component $X_0 \subseteq X$ (so X_0 is a tree) corresponding to a splitting (I_0, F_0) with $|I_0| \geqq |I|$ and, if $|X_0| > 1$, a maximal element (x', x'') in I_0 with $x', x'' \in X_0$ such that the only possible minimal elements of F_0 are* min (x') *and* min (x'').

DEFINITION. The component X_0 above is a *minimal tree.*

LEMMA 2. *If some closed component of T is a tree, then $k \geqq 2n-4$.*

Proof. From Proposition 3, we have a minimal tree X with canonical splitting (I, F) such that $|I| \geqq n-2$. Either $|X| = 1$ and Lemma 1 applies, or we have elements $x', x'' \in X$ such that min (x') and min (x'') are the only possible minimal elements of F. Since no node except x' and x'' can have received information from both x' and x'' during the calls of I, all nodes other than x' and x'' are members of some call in F. By Proposition 3, $\{\min(u): u \neq x', x''\}$ are distinct elements of F, so $|F| \geqq n-2$.

LEMMA 3. *If $k = 2n-4$, and some closed component X corresponding to a splitting (I, F) is a tree, then $|I| = |F| = n-2$. If X is a minimal tree, then F has precisely two minimal elements, and every element of F is* min (u) *for some u.*

Proof. By Proposition 3, construct a minimal tree X_0 in X and its canonical splitting (I', F'). However, while proving Lemma 2, we showed that $|I'|, |F'| \geqq n-2$, so we must have $|I'| = |F'| = n-2$. By construction, $|I| \leqq |I'|$, and $|I| \geqq n-2$ since (I, F) generates components, so $|I| = |F| = n-2$. If X is a minimal tree, $\{\min(u): u \neq x', x''\}$ already gives all $n-2$ elements of F. Because $|I| = n-2$, $X \neq U$, and x', x'' must receive future calls, so min (x') and min (x'') must exist and be min (u) for some $u \neq x', x''$. This implies that they are distinct, and Proposition 1 shows they are exactly the minimal elements of F.

If $k = 2n-4$ and X is a minimal tree with canonical splitting (I, F), Lemma 3 shows that X and F determine x' and x''. Note that $G(F)$ must consist of two trees with all information flowing outward from x' and x''.

This lemma demonstrates why minimal trees are called "minimal." Suppose that a minimal tree X contained a smaller tree component. This could only be if X had a

maximal element other than $\{x', x''\}$. But, this would allow us to construct a splitting that contradicts Lemma 3: Raise this other maximal element into F and lower min (x') and min (x'') into I. Since X is closed, all these calls are incomparable, and the part of X left without $\{x', x''\}$ is a closed tree. But, we have now made $|I| = n-1$, so the canonical splitting for this tree must have $|I| \geqq n-1$. Thus, in a minimal tree, the calls of I transfer information inward along the tree into x' and x''. This fact about minimal trees is not needed for our proofs, but is a useful tool; see, e.g., [6].

5. Completion of the proof of Theorem 1. We begin with a construction.

PROPOSITION 4. *Suppose (I, F) is a splitting with $|I| \geqq n-2$ such that at least two components of $G(I)$ are trees. Then one of these components is closed or both contain closed components (which must be trees).*

Proof. Call the components X and Y. If either is closed we are done, so we may assume by Proposition 2 that F has a minimal element y which joins two nodes of Y. X is not a point, so I contains a maximal element $x \in X$.

Interchange x and y, that is, raise x into F and lower y into I, to form a splitting (I', F'). Thus, $|I'| = |I|$, and in $G(I')$ the points of X fall into two components X_0 and X_1, while the nodes of Y no longer span a tree in $G(I)$.

Each X_i is a tree, so we may apply this construction inductively to produce a closed component inside X. Reversing the roles of X and Y produces a closed component inside Y.

LEMMA 4. *There is a closed component of T which is a tree.*

Proof. Let (I, F) be any splitting with $|I| = n-2$. Counting edges of $G(I)$ shows that at least two components of $G(I)$ are trees. Proposition 4 gives the desired result.

Proof of Theorem 1. Lemma 2 and Lemma 4 prove Theorem 1.

Proposition 4, together with Lemma 3, gives additional insight into the structure of pooling systems with $k = 2n-4$. The remainder of this section is devoted to such results which are not needed for the proof of Theorem 2, but are of independent interest. So we now concentrate on T which are pooling with $k = 2n-4$. We let (I, F) be a splitting satisfying the hypothesis of Proposition 4.

If one of the tree components, say X, is closed, then Lemma 3 implies that $|I| = n-2$ and every minimal element of F has one end in X. Thus, all components are closed. Now applying the same analysis to Y, we find that the minimal elements of F must link X and Y. There can be no further tree components, so every other component must have the same number of edges as nodes. If u is a minimal element of F, then one component of $G(I \cup u)$ is a tree containing X and Y. Again, Lemma 3 tells us that this component is not closed, so adding an element of its closure gives a splitting (I', F') with $|I'| = n$ and all components of $G(I')$ having the same number of edges as nodes. Call such a splitting "balanced."

Now suppose that the tree components of $G(I)$ are not closed. The construction of Proposition 4 does not change $|I|$, nor does it alter any component disjoint from the trees X and Y. When we are finished, we have a closed tree component. Thus, $|I| = n-2$ and any component other than X or Y is closed and has the same number of edges as nodes. If we add a minimal element of F joining two nodes of X and a minimal element of F joining two nodes of Y to the given I, we will obtain a balanced splitting (I', F').

We now show that a balanced splitting has at most two components and that these components are closed. A single component arises only from a splitting (I, F) for which $G(I)$ has only two components, both closed trees. We may then limit ourselves to the case in which $G(I')$ has more than one component. Remove any maximal element from I'. The resulting graph has one tree component, and the total number of components is either the same or increased by one. By Lemma 3, the tree cannot be closed and hence,

by the corollary to Proposition 2, cannot consist of a single point. Now remove a maximal element of this tree to get a splitting satisfying the hypothesis of Proposition 4. As we have already noted, all components other than the trees are closed, so all components of $G(I')$ except the one we dissected are closed. But we could have chosen any component, so all components are closed.

If $G(I')$ has more than two components, add a minimal element of F'. This connects only two components. Dissecting any other components leads to a splitting satisfying the hypothesis of Proposition 4 for which $|I| > n-2$, which we have seen contradicts Lemma 3. Thus, $G(I')$ can have only two components. Also, if the removal of a maximal element of I' were to give a graph with three components, we could add a link to the tree component to give a new balanced splitting with three components. Thus, this possibility is also ruled out. Finally, removing a maximal link from the tree component gives a splitting with $|I| = n-2$ and two tree components. If these trees were not closed, we could construct a balanced splitting with three components by closing them. Thus:

THEOREM 3. *If T is pooling with $k = 2n-4$ and (I, F) is a splitting with $|I| \geqq n-2$ with at least two tree components in $G(I)$, then $|I| = n-2$ and one of these cases holds:*

Case I. *The two trees are closed, and they are the only components.*

Case II. *The two trees are closed and there is one other component. If one removes a maximal link from this component, one gets a single tree; and if one removes a maximal link from this, then one gets two closed trees. There is an $I'' \subseteq I$ such that $G(I'')$ is the union of four closed trees.*

Case III. *The trees are not closed. Now there can be no further components. Removal of a maximal link from each of the components again gives closed trees.*

6. Blocks. Let T be a pooling system with $k = 2n-4$ and (I, F) a splitting with some minimal tree component X. This gives, from Lemma 3, elements $x', x'' \in X$ such that $F = \{\min(u): u \neq x', x''\}$. This X will be fixed for the rest of the section.

PROPOSITION 5. *With these assumptions, if $u \in U - X$, then there are elements u', u'' such that $u \to u'$ and $u \to u''$ in I and* $\min(u') = \{u', x'\}$ *and* $\min(u'') = \{u'', x''\}$.

Proof. Consider the path $u = u_0, u_1, \cdots, u_{i-1}, u_i = x'$ proving $u \to x'$ in T. Since u and x belong to different components of $G(I)$, some step must be in F, and hence, all steps from that point on must be in F. The last step, $\{u_{i-1}, x'\}$ must be $\min(u_{i-1})$, but then $\{u_{i-2}, u_{i-1}\}$ cannot belong to F. The element u_{i-1} is the desired u'. We find u'' the same way, starting from $u \to x''$.

Note that u' and u'' must lie on the same component of $G(I)$ as u does. From a count of edges of $G(I)$, there is a component Y of $G(I)$, different from X, which is a tree. The component Y is an example of a "block."

DEFINITION. A tree component Y is a *block* if for each $y \in Y$, there are $y', y'' \in Y$ such that $y \to y', y''$ in Y and $\min(y') = \{x', y'\}$ and $\min(y'') = \{x'', y''\}$.

Knowing that blocks exist, we will construct "minimal blocks" by inductively reducing the size of a block until certain properties hold. If $t = \{y_0, y_1\}$ is a maximal element in Y, then any sequence of calls that proves $u \to v$ in Y must prove $u \to v$ in $Y - t$ or else the last step is t. In the latter case, $v = y_0$ and $u \to y_1$ in $Y - t$ or $v = y_1$ and $u \to y_0$ in $Y - t$. $G(Y - t)$ has two components Y_0, Y_1 with $y_i \in Y_i (i = 0, 1)$.

If $\min(y_1) \neq \{y_1, x'\}$ or $\{y_1, x''\}$, then y_1 can never be a y' or y''. From this it follows that Y_0 is a block. Similarly, if $\min(y_0) \neq \{y_0, x'\}$ or $\{y_0, x''\}$, then Y_1 is a block. In either case, we get a smaller block.

Now suppose that $\min(y_0)$ and $\min(y_1)$ both involve the same element of X, say x'. If $y \in Y_0$, then $y \to y''$ in $Y - t$ so $y'' \in Y_0$. Either $y' \in Y_0$, in which case we also have $y \to y'$

in $Y-t$; or $y'=y_1$. In the latter case, $y \to y_0$ in $Y-t$, so we could use y_0 for y' instead. In either case, Y_0 is a smaller block than Y. (In fact, Y_1 is also.)

LEMMA 5. *Minimal blocks exist, and if Y is a minimal block with y_0, y_1 a maximal edge, then either $(x'-x''-y_0-y_1-x')$ or $(x'-x''-y_1-y_0-x')$ is a four-cycle in $G(T)$.*

Proof. Induction on the above construction gives a minimal block. The only minimal blocks are where y_0 and y_1 are adjacent to different elements of $\{x', x''\}$.

This completes the proof of Theorem 2.

Acknowledgment. The present form of the paper owes much to the careful reading of earlier versions by the referee. The author would also like to express special thanks to Dale Worley for editorial work in the preparation of the final manuscript.

REFERENCES

[1] B. BAKER AND R. SHOSTAK, *Gossips and telephones*, Discrete Math., 2 (1972), pp. 191–193.

[2] R. GUY, *Monthly Research Problems*, 1969–75, Amer. Math. Monthly, 82 (1975), pp. 995–1004 (this problem discussed on p. 1001).

[3] A. HAJNAL, E. C. MILNER AND E. SZEMERÉDI, *A cure for the telephone disease*, Canad. Math. Bull., 15 (1976), pp. 447–450.

[4] F. HARARY AND A. J. SCHWENK, *The communication problem on graphs and digraphs*, J. Franklin Inst., 297 (1974), pp. 491–495.

[5] ———, *Efficiency of dissemination of information in one-way and two-way communictation networks*, Behavioral Science, 19 (1974), pp. 133–135.

[6] D. J. KLEITMAN AND J. B. SHEARER, *Further gossip problems*, Discrete Math., 30 (1980), pp. 191–193.

[7] R. TIJDEMAN, *On a telephone problem*, Nieuw Arch. Wisk., 3 (1971), pp. 188–192.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 1, March 1981

0196-5212/81/0201-0003 $01.00/0

ON THE SENSITIVITY OF THE GRAVITY MODEL*

TOMMY ELFVING†

Abstract. The gravity model is perhaps the most widely used mathematical method for predicting travel between subareas in an urban region. In this paper, we will consider the Evans-Kirby version of the model [Transpn. Res., 8 (1974), pp. 105–122]. A formula is derived which shows the sensitivity of the model to errors in data from the prediction year, say ten years in the future, and to errors in data from the base year.

1. Introduction. A study of transportation in a region usually involves partitioning the region into zones. In the trip generation step of transportation planning, the number of future trips leaving each zone and the number of trips arriving in each zone are estimated. Given these numbers, the next step is to predict how trips will be distributed among zones, on a pairwise basis. This phase is called the trip generation step. The gravity model is probably the most widely used method for making this distribution. The model is based on the assumption that trips tend to be distributed in inverse proportion to the distance (or more generally, to the cost of travel) between zones. We refer to a recent paper, including a comprehensive bibliography, by Erlander [4], for a survey of different aspects of the model. Stewart in [9] surveys several traffic assignment and distribution models, including the gravity model.

In this paper, we will analyze the sensitivity of the gravity model, calibrated with the procedure proposed by Evans and Kirby [5], to errors from the trip generation step. In [6], Jensen and Stewart have studied the same subject. The purpose of this paper is to simplify and extend their results. In the following section, we will introduce notation and formulate the problem. The sensitivity analysis is carried out in § 3, and is summarized in (3.9) and the error bounds (3.10) and (3.11).

2. The problem. Let $S=\{k_1, k_2, \cdots, k_I\}$ be the set of all zones where a trip may begin and $T=\{m_1, m_2, \cdots, m_J\}$ the set of all zones where a trip may end. Denote by x_{ij} the number of trips from zone k_i to zone m_j. If it is desired to avoid predicting intrazonal trips, any unknown x_{ij} such that $k_i = m_j$ should be excluded from the model. Let $\{g_i\}_1^I$ and $\{a_j\}_1^J$ be estimates of the number of trips departing and arriving in each zone. Then the transportation constraints become

$$\begin{aligned} &\sum_{j=1}^{J} x_{ij} = g_i, \qquad i=1,2,\cdots,I, \\ &\sum_{i=1}^{I} x_{ij} = a_j, \qquad j=1,2,\cdots,J. \end{aligned} \tag{2.1}$$

Define $x=(x_{11}, x_{12}, \cdots, x_{1J}, x_{21}, \cdots, x_{IJ})^T$, $b'=(g_1, g_2, \cdots, g_I)^T$ and $b''=(a_1, a_2, \cdots, a_J)^T$. We then write (2.1) in compact form,

$$Ax=b, \quad \text{with } A^T=(A_1^T, A_2^T) \quad \text{and} \quad b^T=(b'^T, b''^T). \tag{2.2}$$

The following example, where $I=J=3$, illustrates the nonzero structure of the matrix,

$$A_1=\begin{bmatrix} 1 & 1 & 1 & & & & & & \\ & & & 1 & 1 & 1 & & & \\ & & & & & & 1 & 1 & 1 \end{bmatrix}, \qquad A_2=\begin{bmatrix} 1 & & & 1 & & & 1 & & \\ & 1 & & & 1 & & & 1 & \\ & & 1 & & & 1 & & & 1 \end{bmatrix}.$$

* Received by the editors October 3, 1979 and in revised form July 12, 1980.
† National Defense Research Institute, Box 1165, S-58 111 Linköping, Sweden.

In the gravity model, the predicted number of trips is assumed to have the form

$$x_{ij} = p_i f(c_{ij}) q_j,$$

where the positive numbers p_i, q_j are unknowns. The function f is called the deterrence function and measures the effect of the cost on the amount of travel from zone k_i to zone m_j. The word cost may be interpreted very liberally, to include monetary cost, travel discomfort, etc. The name gravity model comes from the possible choice $f = 1/c(i, j)$, where here $c(i, j)$ is the squared distance between zone k_i and zone m_j.

We will now introduce some further notation. Following [6], we distinguish between data from the calibration year (or base year), which will be denoted by superscript 0 (for vectors subscript c), and data from the prediction year, denoted by superscript 1 (for vectors subscript p). Let $I_1, I_2, \cdots, I_K$ be K finite disjoint intervals on the nonnegative real line. Assume that the following base year data are known:

(i) $c_c = (c_{11}^{(0)}, c_{12}^{(0)}, \cdots, c_{1J}^{(0)}, c_{21}^{(0)}, \cdots, c_{IJ}^{(0)})^T$, where $c_{ij}^{(0)}$ equals the cost of travel from zone k_i to zone m_j in the calibration year.

(ii) $b_c' = (g_1^{(0)}, g_2^{(0)}, \cdots, g_I^{(0)})^T$, where $g_i^{(0)}$ equals the number of trips departing from zone k_i in the calibration year.

(iii) $b_c'' = (a_1^{(0)}, a_2^{(0)}, \cdots, a_J^{(0)})^T$, where $a_j^{(0)}$ equals the number of trips arriving in zone m_j in the calibration year.

(iv) $b_c''' = (s_1^{(0)}, s_2^{(0)}, \cdots, s_K^{(0)})^T$, where $s_k^{(0)}$ equals the number of trips with $c_{ij}^{(0)} \in I_k$ in the calibration year.

We also introduce the notation:

(v) $b_c^T = (b_c'^T, b_c''^T, b_c'''^T)$.

(vi) $x_c = (x_{11}^{(0)}, x_{12}^{(0)}, \cdots, x_{1J}^{(0)}, x_{21}^{(0)}, \cdots, x_{IJ}^{(0)})^T$.

The first step is the calibration of the model in the base year. This means finding a function $f(c_{ij})$ such that the frequency distribution of the costs in the base year matches the distribution produced by the model. This implies that the following calibration condition shall be satisfied:

$$\sum_{(i,j)} x_{ij}^{(0)} = s_k^{(0)}, \qquad 1 \leqq k \leqq K. \tag{2.3}$$

The summation is over all (i, j) such that $c_{ij}^{(0)} \in I_k$. The following compact notation will be used for this condition:

$$A_3^{(0)} x_c = b_c'''. \tag{2.4}$$

We illustrate the nonzero structure of the matrix $A_3^{(0)}$ by the following example. Let $I = J = K = 3$ and take $I_1 = (0.1, 1)$, $I_2 = (1.1, 2)$ and $I_3 = (2.1, 3)$. Assume that $c_c = (2.3, 1.6, 0.7, 2.9, 2.2, 1.5, 0.1, 1.7, 2.5)^T$, and hence that $s_1^{(0)} = 2$, $s_2^{(0)} = 3$ and $s_3^{(0)} = 4$. The matrix $A_3^{(0)}$ then becomes

$$A_3^{(0)} = \begin{bmatrix} & 1 & & & 1 & \\ 1 & & & 1 & & 1 \\ & & 1 & 1 & & \end{bmatrix}.$$

The calibration problem has been solved by Evans and Kirby [5] by assuming that f is piecewise constant,

$$f(c_{ij}) = \begin{cases} r_k^{(0)} & \text{if } c_{ij} \in I_k, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

The calibration step can be summarized as follows. Find $p_i^{(0)}$, $q_j^{(0)}$ and $r_k^{(0)}$, given estimates of c_c and b_c such that the following relations are fulfilled:

$$x_{ij}^{(0)} = p_i^{(0)} f(c_{ij}^{(0)}) q_j^{(0)}, \tag{2.6a}$$

$$A_c x_c = b_c, \qquad x_c \geqq 0, \quad \text{where } A_c^T = (A_1^T, A_2^T, A_3^{(0)T}). \tag{2.6b}$$

This process can also be viewed as a rule for choosing the so-called socioeconomic factors in such a way as to match exactly any given histogram of trip costs in the calibration year [8] (see also [9, Chapt. 1]).

The second step of the process is called the prediction step. Here, we assume known the values of $c_{ij}^{(1)}$, $a_j^{(1)}$ and $g_i^{(1)}$ from the prediction year. The vectors c_p, b_p', b_p'' and x_p are defined analogously to (i), (ii), (iii) and (vi) above. The prediction step can be summarized as follows. Find $p_i^{(1)}$ and $q_j^{(1)}$, given estimates of c_p, b_p', b_p'' and $\{r_k^{(0)}\}$, such that

$$x_{ij}^{(1)} = p_i^{(1)} f(c_{ij}^{(1)}) q_j^{(1)}, \tag{2.7a}$$

$$A_p x_p = b_p, \qquad x_p \geqq 0, \quad \text{where } A_p^T = (A_1^T, A_2^T) \text{ and } b^T = (b_p'^T, b_p''^T). \tag{2.7b}$$

We remark that, given a deterrence function $f(c_{ij})$, the vectors x_p and x_c (occurring in (2.6) and (2.7) respectively) are always uniquely determined. However, the factors p_i, q_j and, in the calibration case, $r_k^{(0)}$, are not unique unless the matrices of the side conditions have full rank; see, e.g., [3]. As observed in [2], this creates a problem if rank $(A_c) < I + J + K$, since then the function $f(c_{ij})$ in (2.5) is not unique. The matrix A_c has, in general, two redundant rows [6]. We will assume, to avoid this difficulty, that the redundant rows have been deleted in A_c. We will also mention the well-known fact that the matrix A_p has one redundant row. Whether this row is deleted or not does not influence the vector x_p. We remark, however, from computational experience, that the rate of convergence of algorithms for solving (2.7) decreases if a full rank matrix is used. In fact, for a special case, it is shown [3], that the so-called balancing method for solving (2.7) converges quadratically if all rows are kept in A_p, but only linearly if one row is deleted.

3. The sensitivity analysis. In this section we will carry out a first-order sensitivity analysis, i.e., neglecting error terms higher than one, of the process described in § 2. Let x be a vector of length n. We will use the notation $\tilde{x}$ for the perturbed value of x. The absolute error, Δx, is defined from the relation $\tilde{x} = x + \Delta x$. For a nonzero x, the relative error is defined as $\delta x = D_x^{-1} \Delta x$, where D_x is a diagonal matrix such that the jth diagonal element equals the jth component of x. To avoid trivial exceptions, we will assume that both x_c and x_p are positive so that relative errors are well defined.

In the calibration step, possible errors originate from b_c and from the generalized costs c_c. For a detailed discussion of possible sources for these errors, see [6]. Consider a specific cost $\tilde{c}_{ij}^{(0)} = c_{ij}^{(0)} + \Delta c_{ij}^{(0)}$. We note that unless $\Delta c_{ij}^{(0)}$ is such that $\tilde{c}_{ij}^{(0)} \in I_{k1}$ and $c_{ij}^{(0)} \in I_{k2}$, $k1 \neq k2$, the error in the cost is irrelevant. If however, $k1 \neq k2$, then two elements are changed in the matrix $A_3^{(0)}$, cf. (2.3), whose nonzero elements equal $+1$. This is, of course, not a small change. Hence, we can expect that a first-order analysis will not cover this case. In fact, Jensen and Stewart give a simple example which shows that the gravity model can predict completely different outcomes due to such errors. We shall in the sequel exclude the case $k1 \neq k2$ in our analysis (both in the calibration and prediction step) and hence ignore errors in the generalized costs. One way to avoid having c_{ij} and $\tilde{c}_{ij}$ fall into different cost intervals is to choose these intervals well separated with respect to the errors [6].

We will now estimate the errors in the deterrence function due to errors in the vector b_c. Consider, therefore, the following perturbed calibration problem, cf. (2.6):

$$\tilde{x}_{ij}^{(0)} = \tilde{p}_i^{(0)} \tilde{f}(c_{ij}^{(0)}) \tilde{q}_j^{(0)}, \tag{3.1a}$$

$$A_c \tilde{x}_c = \tilde{b}_c. \tag{3.1b}$$

By neglecting higher-order terms, we obtain

$$\delta x_{ij}^{(0)} = \delta p_i^{(0)} + \delta f(c_{ij}^{(0)}) + \delta q_j^{(0)}, \tag{3.2a}$$

$$A_c \Delta x_c = \Delta b_c. \tag{3.2b}$$

Define $\delta\beta_c = (\delta p_1^{(0)}, \cdots, \delta p_I^{(0)}, \ \delta q_1^{(0)}, \cdots, \delta q_J^{(0)}, \ \delta r_1^{(0)}, \cdots, \delta r_K^{(0)})^T$. Then it is straightforward to verify that (3.2a) can be written in vector form as

$$\Delta x_c = D_{x_c} A_c^T \delta\beta_c. \tag{3.3}$$

Substituting (3.3) into (3.2b) yields

$$A_c D_{x_c} A_c^T \delta\beta_c = \Delta b_c. \tag{3.4}$$

We now consider the prediction step and note that possible errors come from b_p and $\{r_k^{(0)}\}$. The perturbed prediction problem is

$$\tilde{x}_{ij}^{(1)} = \tilde{p}_i^{(1)} \tilde{u}_{ij}^{(1)} \tilde{q}_j^{(1)} \quad \text{with } \tilde{u}_{ij}^{(1)} = \tilde{f}(c_{ij}^{(1)}), \tag{3.5a}$$

$$A_p \tilde{x}_p = \tilde{b}_p. \tag{3.5b}$$

Note, in contrast to the calibration step, that $\{\tilde{u}_{ij}\}$ is now a *given* estimate of the trip destination table. By neglecting higher-order terms we get

$$\delta x_{ij}^{(1)} = \delta p_i^{(1)} + \delta u_{ij}^{(1)} + \delta q_j^{(1)}, \tag{3.6a}$$

$$A_p \Delta x_p = \Delta b_p. \tag{3.6b}$$

Define $\delta r_c = (\delta r_1^{(0)}, \cdots, \delta r_K^{(0)})^T$, $\delta\beta_p = (\delta r_1^{(1)}, \cdots, \delta p_I^{(1)}, \ \delta q_1^{(1)}, \cdots, \delta q_J^{(1)})^T$ and $\delta u_p = (\delta u_{11}^{(1)}, \delta u_{12}^{(1)}, \cdots, \delta u_{1J}^{(1)}, \delta u_{21}^{(1)}, \cdots, \delta u_{IJ}^{(1)})^T$. Let $A_3^{(1)}$ be defined from c_p in the same way as the matrix $A_3^{(0)}$, (2.4), was defined from the cost vector c_c. Then (3.6a) can be written

$$\Delta x_p = D_{x_p} A_p^T \delta\beta_p + D_{x_p} \delta u_p \quad \text{with } \delta u_p = A_3^{(1)T} \delta r_c. \tag{3.7}$$

From (3.6b) and (3.7) follows

$$A_p D_{x_p} A_p^T \delta\beta_p = \Delta b_p - A_p D_{x_p} \delta u_p. \tag{3.8}$$

We now introduce some further notation. Denote by $D_x^{1/2}$ the Choleski factor of D_x, and by I the identity matrix. Let $\hat{A}_p = A_p D_{x_p}^{1/2}$ and define $\hat{A}_p^+ = \hat{A}_p^T (\hat{A}_p \hat{A}_p^T)^+$, the pseudoinverse of $\hat{A}_p$. From (3.7) and (3.8), it follows after some calculations that

$$\Delta x_p = D_{x_p}^{1/2} \hat{A}_p^+ \Delta b_p + D_{x_p}^{1/2} (I - \hat{A}_p^+ \hat{A}_p)_{x_p}^{1/2} \delta u_p, \tag{3.9}$$

where $P = I - \hat{A}_p^+ \hat{A}_p$ is the orthogonal projector onto the nullspace of $\hat{A}_p$ (and hence, $\|P\| \leqq 1$). Note that although (3.8) does not have a unique solution, the perturbation Δx_p is always unique. Denote by $\kappa(\hat{A}_p) = \|\hat{A}^+\|_2 \cdot \|\hat{A}_p\|_2$ the Euclidean condition number of $\hat{A}_p$. We remind the reader of the relations $\kappa(\hat{A}_p \hat{A}_p^T) = \kappa^2(\hat{A}_p) = \sigma_{\max}^2 / \sigma_{\min}^2$, where $\sigma_{\max}^2$ and $\sigma_{\min}^2$ are the largest and smallest (nonzero) eigenvalues of the matrix $\hat{A}_p \hat{A}_p^T$, respectively, [1, p. 241]. Before we give the bound for the perturbation in x_p, we will

need some minor results. We first note that

$$b_p = A_p x_p = \hat{A}_p D_{x_p}^{1/2} e \quad \text{with } e = (1, 1, \cdots, 1)^T.$$

The following relations between norms are easily verified:

$$\|D_{x_p}^{1/2}\|_2^2 = \|x_p\|_\infty \quad \text{and} \quad \|D_{x_p}^{1/2} e\|_2^2 = \|x_p\|_1.$$

Let $k_p^2 = \|x_p\|_1 / \|x_p\|_\infty$. From (3.9) follows

$$\frac{\|\Delta x_p\|_2}{\|x_p\|_\infty} \leqq k_p \kappa(\hat{A}_p) \frac{\|\Delta b_p\|_2}{\|b_p\|_2} + \|\delta u_p\|_2. \tag{3.10}$$

Let $\hat{A}_c = A_c D_{x_c}^{1/2}$. From (3.4), (3.10) and the inequality $\|\delta u_p\|_2 = \|A_3^{(1)^T} \delta r_c\|_2 \leqq \|A_3^{(1)^T}\|_2 \cdot \|\delta \beta_c\|_2$, we arrive at

$$\frac{\|\Delta x_p\|_2}{\|x_p\|_\infty} \leqq k_p \kappa(\hat{A}_p) \frac{\|\Delta b_p\|_2}{\|b_p\|_2} + k_c \kappa^2(\hat{A}_c) \frac{\|\Delta b_c\|_2}{\|b_c\|_2}, \tag{3.11}$$

where $k_p^2 = \|x_p\|_1 / \|x_p\|_\infty$ and $k_c = \|A_3^{(1)^T}\|_2 \cdot \|A_c x_c\|_2 / \|A_c D_{x_c} A_c^T\|_2$. We remark that the parameter k_c is invariant under scaling of the constraints $A_c x_c = b_c$, the reason being that in the preceding derivations, cf., e.g., (3.3), we used precisely the original definition of A_c.

We now compare the results given here with those obtained by Jensen and Stewart. In order to analyze the effects of possible errors in the costs, a three-dimensional model (with unknowns x_{ijk}) is used in [6]. As discussed earlier, these errors are neglected here. When comparing results, we therefore used the relation $x_{ij} = \sum_k x_{ijk}$. It is then possible to verify that the relations (3.4) and (3.8) correspond to the relations (19) and (23), respectively, in [6]. However, the analysis in [6] is not carried further.

By making this extension, some further conclusions can be drawn. For instance, it is seen from (3.10) that possible perturbations in the initial estimate, $\{u_{ij}^{(1)}\}$, of the trip destination table are not magnified by the condition number $\kappa(\hat{A}_p)$, cf. [6]. For the case when $\kappa(\hat{A}_p)$ and $\kappa(\hat{A}_c)$ are of the same order relation, (3.11) shows that the bound for the final error is less sensitive to errors in data from the prediction year, say ten years in the future, than to errors from the base year.

In a recent report [7], Murchland presents an easily computable first-order formula for $\delta x_{ij}^{(1)}$, due to possible errors in the prediction data $\{u_{ij}^{(1)}\}$, $\{g_i^{(1)}\}$ and $\{a_j^{(1)}\}$. Although no derivation is given, his result is apparently an approximation of the right-hand side of (3.9). It is stated that the formula is valid provided the balancing algorithm for solving problem (2.7), converges rapidly. The approach in [7] does not seem to lead easily to a bound corresponding to (3.10).

We remark finally that another way to carry out this analysis would be to consider the related maximum entropy problem with linear constraints. This approach also conveniently allows for perturbations in the matrix elements. We will return to this sensitivity problem elsewhere.

Acknowledgment. The author is grateful to Professor Neil F. Stewart, Université de Montréal, Canada, for a critical reading of the manuscript.

REFERENCES

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

[2] M. J. L. DAY AND A. F. HAWKINS, *Partial matrices, empirical deterrence functions and ill-defined results*, Traffic Eng. and Control, 20 (1979), pp. 429–433.

[3] T. ELFVING, *On some methods for entropy maximization and matrix scaling*, Linköping University, 1979, Linear Algebra Appl., to appear.
[4] S. ERLANDER, *Optimal Spatial Interaction and the Gravity Model*, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1980.
[5] S. P. EVANS AND H. R. KIRBY, *A three-dimensional Furness procedure for calibrating gravity models*, Transpn. Res., 8 (1974), pp. 105–122.
[6] F. JENSEN AND N. F. STEWART, *A sensitivity analysis of the gravity model*, Infor., 15, (1977), pp. 308–321.
[7] J. D. MURCHLAND, *Perturbation of the biproportional model*, 1978: 12:22 Transport Studies Group, University College, London.
[8] R. B. POTTS AND R. M. OLIVER, *Flows in Transportation Networks*, Academic Press, New York, 1972.
[9] N. F. STEWART, *Notes on the mathematical structure of equilibrium models*, Linköping University, LiTH-MAT-R-79-8, Linköping, Sweden, 1979.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 1, March 1981

0196-5212/81/0201-0004 $01.00/0

SENSITIVE GROWTH ANALYSIS OF MULTIPLICATIVE SYSTEMS I: THE DYNAMIC APPROACH*

URIEL G. ROTHBLUM†

Abstract. Consider a system in which a vectorial input x is transferred, in one time period, into a vectorial output xP, where P is a nonnegative matrix (not necessarily irreducible). We study the asymptotic behavior of the n-period output of such a system. For each i, we define two growth coefficients: the geometric growth rate, say r_i, and the power growth rate, say ν_i. We show that if $x(n)_i$ is the n-period output of the ith coordinate and $r_i \neq 0$, then $r_i^{-n}x(n)_i$ can asymptotically be a Cesaro average approximated by a polynomial. A combinatorial characterization of $\nu_i - 1$ which turns out to be the degree of the approximating polynomial is obtained. We also show that $r_i^{-n}x(n)_i$ has a polynomial periodic asymptotic behavior.

1. Introduction. Consider a multiplicative system in which a nonnegative vectorial input x is transferred, during a single period, into an output vector xP, where P is a square nonnegative matrix. Assume that this system is operating for a number of periods by using the output at the end of each period as the input of the system at the beginning of the next period. There are many examples for systems that operate in this way. In particular, finite state Markov chains have the above structure with the input/output vector representing probability distributions. Also, branching processes (e.g., Harris (1963)), Markov reward processes with exponential utility (e.g., Howard and Matheson (1972)) and age distribution models have this structure. The terms we use for our analysis are motivated by a simple production model of a self-sustained economy in which there are a finite number of commodities. During each time period the bundle of commodities available at the beginning of that period undergoes a production process in which P_{ij} units of the jth commodity are produced from each unit of commodity i. So, each commodity can (under the production process) create a variety of output commodities. We point out that this model is different from the closed Leontief model in which a variety of input commodities are used to produce one unit of an output product. This model is the general von Neumann model with the input matrix being the identity and the output matrix being arbitrary.

If a multiplicative system is operated for n periods, then every input vector $x(0)$ is transferred into $x(n) = x(0)P^n$. In this paper, we study the asymptotic behavior of the n-period output as n becomes large. This asymptotic behavior has been studied extensively under various assumptions on the structure of the transition matrices. It has been shown that if r is the spectral radius of the transition matrix and the system is either indecomposable (see § 3) or satisfies some other structural restrictions, then $r^{-n}x(n)$ has a nonvanishing finite (possibly Cesaro average of order one) limit as $n \to \infty$. For specific examples, see Karlin (1959), (1966), Kemeny and Snell (1960), Jaquette (1975), Harris (1963), Gale (1960), Nikaido (1968), Howard and Matheson (1972) and many others. When $r^{-n}x(n)$ converges, r can be considered to be the growth rate of the system. Unfortunately, this convergence does not always hold. McKenzie (1967), (1971) showed that, qualitatively, the n-period output of some production system might have a polynomial behavior on top of the geometric growth. In this paper, we give a complete

* Received by the editors August 24, 1979 and in revised form May 19, 1980. This research was supported by the National Science Foundation under grant ENG-78-25182. This work also relates to the U.S. Department of the Navy contract ONR 00014-76-C-0085 issued under U.S. Office of Naval Research contract authority NR047-006.

† Yale University, School of Organization and Management, New Haven, Connecticut 06520. This paper was revised while the author was visiting the Faculty of Industrial Engineering and Management at the Technion, Haifa, Israel.

and precise expansion of the n-period output. Of course, examining the ijth element of the nth power of the nonnegative matrix P amounts to considering $[x(0)P^n]_j$ where $x(0)$ is the ith unit vector.

Our analysis uses the expansions of partial sums of matrix powers developed in Rothblum and Veinott (1975) and Rothblum (1980) (summarized in § 2) with the spectral class structure of a nonnegative matrix first observed in Rothblum (1975) (summarized in § 3). We then combine these structural studies in § 4. Specifically, for each commodity[1] i we define the *geometric* and *power growth rates*, respectively, by

$$r_i \equiv \inf\left\{\alpha > 0 \,\middle|\, \lim_{n\to\infty} \alpha^{-n} x(n)_i = 0\right\},$$

and if $r_i \neq 0$,

$$\nu_i \equiv \inf\left\{k = 0, 1, \cdots \,\middle|\, \lim_{n\to\infty} n^{-k} r_i^{-n} x(n)_i = 0\right\}.$$

We study some properties and characterizations of these growth coefficients. In particular, we show that ν_i equals the number of independent nondegenerate subsystems producing each other all of which can produce i and have the maximal geometric growth rate among all subsystems producing i. We then show that for each commodity i, $r_i^{-n}x(n)_i$ is Cesaro-average approximated by a polynomial of order $\nu_i - 1$, and this sequence has also a periodic approximation by polynomials whose degree does not exceed $\nu_i - 1$ (see § 4 for details). Specifically, we show the existence of integers τ_i and q_i and of (computable) vector-polynomials $\phi(\cdot), \phi_0(\cdot), \cdots, \phi_{q_i-1}(\cdot)$ of degree $d, d_0, \cdots, d_{q-1}$, respectively, such that

$$\lim_{n\to\infty} \{r_i^{-n} x(n)_i - \phi(n)\} = 0 \qquad (C, \tau_i), \tag{1.1}$$

and

$$\lim_{m\to\infty} \{r_i^{mq_i+j} x(mq_i + t)_i - \phi_j(m)\} = 0, \qquad j = 0, \cdots, q_{i-1}, \tag{1.2}$$

where (C, τ_i) stands for the Cesaro average of order τ_i (see § 2 for a precise definition) and $d = \max_{0\leqq j\leqq q_{i-1}} d_j = \nu_i - 1$. We emphasize that $x(n)_i$ is normalized by a term which depends on i. If $x(n)_i$ is normalized, r^{-n} where $r > r_i$, we trivially would have that $\lim_{n\to\infty} r^{-n}x(n)_i = 0$ and little information is gained on $x(n)_i$.

2. Notational conventions and preliminary results. We say that a (real) matrix A is *nonnegative* (resp., *positive*) written $A \geqq 0$ (resp., $A \gg 0$) if all its coordinates are nonnegative (resp., positive). We say that A is *semipositive*, written $A > 0$, if $A \geqq 0$ and $A \neq 0$. We write $A \geqq$ (resp., $\gg$ or $>$) B if $A - B \geqq$ (resp., $\gg$ or $>$) 0. Similar definitions apply to vectors.

The real line will be denoted R. Let $B \in R^{S\times S}$ and let $J, K \subseteq \{1, \cdots, S\}$. Then by $B_{JK} \in R^{|J|\times|K|}$ we denote the corresponding submatrix of rows and columns of B.[2] Let $B_J \equiv B_{JJ}$. For $x \in R^S$ and $J \subseteq \{1, \cdots, S\}$ we denote by $x_J \in R^{|J|}$ the corresponding subvector of x. If $I = \{i\}$ and $J = \{j\}$ we put (as usual) $B_{ij} \equiv B_{IJ}$ and $x_j \equiv x_J$. Superscripts of matrices will stand for powers, whereas superscripts of vectors will be used for enumeration. In particular, $B_{IJ}^n \equiv (B^n)_{IJ}$.

[1] We refer to the indices of the input-output vectors as commodities.

[2] For a finite set L, $|L|$ denotes the number of elements in L. For a complex number λ, $|\lambda|$ denotes the absolute value of λ. No confusion should occur.

For a given sequence $\{a_n | n = 0, 1, \cdots\}$, let $\{a_n^{(m)} | n = 0, 1, \cdots\}$ be the sequence of m-order averages of $\{a_n\}$; i.e., for $n = 0, 1, \cdots$,

$$a_n^{(0)} = a_n,$$

and for every positive integer m

$$a_n^{(m)} = (n+1)^{-1} \sum_{i=0}^{n} a_i^{(m-1)}.$$

We say that a is the (C, m) (Cesaro average of order m) limit of $\{a_n\}$, written $\lim_{n\to\infty} a_n = a(C, m)$, if $\lim_{n\to\infty} a_n^{(m)} = a$.[3] It is known that if $\lim_{n\to\infty} a_n = a(C, m)$ and $k \geqq m \geqq 0$, then $\lim_{n\to\infty} a_n = a(C, k)$ (see Hardy (1949, p. 100)).

We next summarize a number of definitions of spectral concepts which we need to summarize results concerning the asymptotic behavior of powers of matrices obtained in Rothblum and Veinott (1975) and Rothblum (1980).

Let P be a square matrix, λ a complex number and $Q \equiv P - \lambda I$. The *index* of λ for P, denoted $\nu_\lambda(P)$, is the smallest nonnegative integer n such that the null spaces of Q^n and Q^{n+1} coincide. The *coindex* of $\lambda \neq 0$ for P, denoted $\tau_\lambda(P)$, is defined by

$$\tau_\lambda(P) \equiv \max \{\nu_\mu(P) | \mu \neq \lambda, |\mu| = |\lambda|\}.$$

It is known (e.g., Halmos (1958, p. 113)) that there exists a unique projection whose range is the null space of Q^ν and whose null space is the range of Q^ν, where $\nu \equiv \nu_\lambda(P)$. This projection is denoted $E_\lambda(P)$ and is called the *eigenprojection* of P at λ. We remark that λ is an eigenvalue of P if and only if $\nu_\lambda(P) > 0$ or equivalently $E_\lambda(P) \neq 0$. In this case, $\nu_\lambda(P)$ is the size of the largest block in the Jordan form corresponding to the eigenvalue λ. Also observe that if P is an $S \times S$ matrix, then $\nu_\lambda(P) \leqq S$.

Let P be square. Then $\sigma(P)$ will denote the *spectrum* of P and $r(P)$ will stand for its spectral radius; i.e., $r(P) \equiv \max \{|\lambda| \, | \, \lambda \in \sigma(P)\}$. It was shown in Rothblum and Veinott (1975) and Rothblum (1980, Theorem 4.2) that for every square matrix P with $r \equiv r(P) \neq 0$,

$$\lim_{n\to\infty} \left\{ r^{-n} P^n - \sum_{j=0}^{\nu-1} \binom{n}{j} Q^j E \right\} = 0 \qquad (C, \tau), \tag{2.1}$$

where $\nu = \nu_r(P)$, $\tau = \tau_r(P)$, $Q = r^{-1}P - I$ and $E = E_r(P)$. It is known that if P is nonnegative then $\tau \leqq \nu$ (e.g., Vandergraft (1968)) and if $r = 0$ then P is nilpotent and $\nu_r(P) = \min \{k | P^k = 0\}$ (e.g., Rothblum (1981)). Moreover, it is also shown in Rothblum (1980) that for some integer q and computable matrix-polynomials $\psi_0, \cdots, \psi_{q-1}$,

$$\lim_{m\to\infty} \{r^{-mq+t} P^{mq+t} - \psi_t(m)\} = 0, \qquad t = 0, \cdots, q-1, \tag{2.2}$$

and

$$\max_{0 \leqq t \leqq q-1} \deg(\psi_t) = \nu - 1.$$

We remark that the (C, τ)-average is needed in order to smoothen fluctuations of order $n^{\tau-1}$ of the sequence whose limit is considered in (2.1). Combinatorial bounds on the order of these fluctuations, namely τ, for nonnegative matrices were obtained in Rothblum (1980, App. C). These bounds demonstrate that, typically, τ is considerably less than ν.

[3] The definition given here is actually that of Hölder limits which are known to be equal to the Cesaro limits (e.g., Hardy (1949, p. 103)).

3. The structure of multiplicative systems. In this section, we introduce a few concepts concerning the structure of multiplicative systems. Most concepts are taken from the theory of nonnegative matrices (e.g., Rothblum (1975)).

Formally, a *multiplicative system* is a triplet $(S, x(0), P)$, where S is a positive integer, $x(0)$ is a $1 \times S$ nonnegative row vector and P is an $S \times S$ nonnegative matrix. We will sometimes omit the term multiplicative and refer to a *system*. The integers $1, \cdots, S$ are called the *commodities* of the system, the matrix P is called the *transition matrix*, and the vector $x(0)$ the *initial input*. The *n-period output* of a multiplicative system $(S, x(0), P)$ is defined as $x(n) = x(0)P^n$, where $n = 0, 1, \cdots$. The *subsystem* of $(S, x(0), P)$ corresponding to $\varnothing \neq K \subseteq \{1, \cdots, S\}$ is the system $(|K|, x(0)_K, P_K)$. The n-period output of this subsystem is $x^K(n) \equiv x(0)_K P_K^n$, where $n = 0, 1, \cdots$. We will usually identify the commodities of a subsystem with the corresponding commodities of the original system.

Let $(S, x(0), P)$ be a given multiplicative system. We say that commodity i *produces* commodity j, or commodity j is *produced by* commodity i, if for some integer $n \geqq 0$, $(P^n)_{ij} > 0$. Two commodities i and j, each producing the other, are said to *communicate*. It is known (e.g., Karlin (1966, p. 42)) that the communication relation is an equivalence relation. Hence, we may partition the totality of commodities into equivalence classes. The commodities in an equivalence class are those which communicate with each other. In the sequel, a *class* will always mean a nonempty equivalence class of communicating commodities. We will use the concept that a class produces (resp., is produced) if this is so for some, or equivalently every, commodity in that class. The classes are partially ordered by the production relation.

We say that our transition matrix is *indecomposable* (sometimes called *irreducible*) if the equivalence relation induces only one class, i.e., if all commodities communicate. It is known that P is indecomposable if and only if $(I+P)^{S-1} \gg 0$ (e.g., Varga (1962, p. 26)). We say that a multiplicative system is *indecomposable* if this is so for its transition matrix.

Let $(S, x(0), P)$ be a multiplicative system. We say that a subset of commodities $\varnothing \neq K \subseteq \{1, \cdots, S\}$ is *essential* (resp., *final*) is no commodity in $\{1, \cdots, S\} \backslash K$ produces (resp., is produced by) any commodity in K. Of course, every system has at least one essential and one final class. Moreover, every essential (resp., final) set of commodities includes at least one essential (resp., final) class. The following lemma is an immediate result of the definition of an essential set.

LEMMA 3.1. *Let K be an essential set. Then for $n = 0, 1, \cdots$, $x(n)_K = x^K(n)$.*

A class J of commodities is called *basic* if $r(P_J) = r(P)$. A class J is called *nonbasic* if it is not basic, i.e., if and only if $r(P_J) < r(P)$ (cf. Varga (1962, p. 30)). Note that the definition of basic class depends on P and not only on P_J; i.e., even if two classes have the same "internal structure" it is possible that when viewed in different systems one is basic and the other is not. However, if K is a union of classes with $r(P_K) = r(P)$, then a class is basic in the subsystem corresponding to K if and only if it is basic in P. Of course, every system has at least one basic class.

Let P be a square nonnegative matrix. A *chain* of classes is a collection of classes such that each class in the collection either produces or is produced by each of the other classes in the collection. A chain of classes with essential class J and final[4] class K is called a *chain from J to K*. The *length* of a chain is the number of basic classes it contains. The *height* of a basic class is the length of the longest chain of classes in which that class is final.

[4] Here essential and final are with respect to the subsystem corresponding to the union of the classes in the chain.

We next illustrate the above definitions by an example. Let our transition matrix be

$$P = \begin{pmatrix} 1 & 0 & 8 & 0 & 0 & 8 & 0 & 0 & 0 \\ & 3 & 6 & 7 & 0 & 0 & 0 & 0 & 0 \\ & & 2 & 0 & 1 & 0 & 0 & 0 & 0 \\ & & & 3 & 3 & 0 & 4 & 0 & 0 \\ & & & & 3 & 5 & 0 & 0 & 0 \\ & & & & & 3 & 0 & 0 & 0 \\ & & & & & & 0 & 2 & 0 \\ & & & & & & & 3 & 3 \\ & & & & & & & & 1 \end{pmatrix}.$$

Obviously, every class consists of a single state. This is done for simplicity only. The basic classes are {2}, {4}, {5}, {6} and {8} and the nonbasic classes are {1}, {3}, {7} and {9}. The production relation between classes can be represented by a directed graph, as done in Fig. 1, where circles are used for basic classes and squares stand for nonbasic

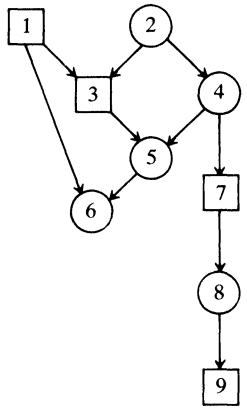


FIG. 1

classes. It is easily seen from Fig. 1 that the essential classes are {1} and {2} and the final classes are {6} and {9}. The height of the basic classes is given by

Basic Class	Height
2	1
4	2
5	3
6	4
8	3

The concepts introduced in this section with respect to the system $(S, x(0), P)$ will sometimes be used with respect to the matrix P; e.g., we will refer to a basic class of a nonnegative matrix P.

4. Asymptotic behavior of the n-period output and growth rate analysis. It is clear from the results summarized in § 2 (see (2.1) and (2.2)) that the asymptotic behavior of the n-period output of a multiplicative system $(S, x(0), P)$ with $r \equiv r(P) \neq 0$ has the

asymptotic expansion given by

$$\lim_{n\to\infty}\left\{r^{-n}x(n)-\sum_{j=0}^{\nu-1}\binom{n}{j}v^j\right\}=0 \qquad (C,\tau), \tag{4.1}$$

and

$$\lim_{m\to\infty}\{r^{-mq+t}x(mq+t)-\phi_t(m)\}, \qquad t=0,\cdots,q-1, \tag{4.2}$$

where $v^j\equiv x(0)Q^jE(j=0,\cdots,\nu-1)$, $\nu\equiv\nu_r(P)$, $\tau\equiv\tau_r(P)$, $Q\equiv r^{-1}P-I$, $E\equiv E_r(P)$, q is some positive integer and $\phi_0,\cdots,\phi_{q-1}$ are (computable) vector polynomials. We remark that since $\tau\leqq\nu$, (C,τ) can be replaced by (C,ν). This expansion shows that once we normalize geometrically by r^n, the n-period output has the corresponding asymptotic polynomial behavior. Previous analysis of models with restrictive structure (e.g., stochastic, indecomposable and age distribution systems) guaranteed that $\nu=1$. In this case, (4.1) shows that $r^{-n}x(n)$ has a $(C,1)$ limit (which equals $E_r(P)$). However, in general, we see that there is a polynomial growth on top of the geometric growth. We remark that an efficient method to compute $E_r(P)$ is described in Rothblum (1976). Thus, one can compute the coefficients $v^j, j=0,\cdots,\nu-1$ explicitly.

The expansion given in (4.1) is obviously of no interest if one studies the growth of a particular commodity i for which it happens that $v_i^j=0$ for $j=0,\cdots,\nu-1$. For this reason, it follows that when a particular commodity is concerned, the expansion (4.1) is not necessarily helpful. For example, if

$$P=\begin{pmatrix}2&2&0\\0&2&0\\0&0&1\end{pmatrix} \quad\text{and}\quad x(0)^T=\begin{pmatrix}1\\1\\1\end{pmatrix},$$

then $r=2$, and

$$r^{-n}x(n)^T=\begin{pmatrix}n+1\\1\\0.5^n\end{pmatrix}.$$

The expansion given in (4.1) tells us that

$$\lim_{n\to\infty}2^{-n}x(n)^T-\begin{pmatrix}1\\1\\0\end{pmatrix}-n\begin{pmatrix}1\\0\\0\end{pmatrix}=0.$$

No information about the n-period output of the third commodity is obtained, except that $2^{-n}x(n)_3\to 0$ as $n\to\infty$. It is also clear that when the second commodity is concerned, one does not need the second term of the expansion. In order to obtain a more sensitive analysis, we next introduce two growth coefficients for each commodity.

For each commodity $i=1,\cdots,S$ define the *geometric growth rate* of i by

$$r_i=\inf\left\{\alpha>0\,\middle|\,\lim_{n\to\infty}\alpha^{-n}x(n)_i=0\right\}. \tag{4.3}$$

If $r_i\neq 0$, the *power growth rate* of i is defined by

$$\nu_i=\min\left\{k=0,1,\cdots\,\middle|\,\lim_{n\to\infty}n^{-k}r_i^{-n}x(n)_i=0\right\}. \tag{4.4}$$

If $r_i=0$, define the *power growth rate* of i by

$$\nu_i=\min\{k=0,1,\cdots|x(n)_i=0 \text{ for all } n=k,k+1,\cdots\}. \tag{4.5}$$

It follows from (4.1) and the discussion thereafter that

(4.6) $$r_i \leqq r(P) \quad \text{for every commodity } i,$$

and

(4.7) $$\nu_i \leqq \nu_{r(P)}(P) \quad \text{for every commodity } i \text{ with } r_i = r(P).$$

We will later show (Theorems 4.3 and 4.4) that the minima in (4.1) and (4.2) are never taken over an empty set.

We say that commodity *i dominates j* if either $r_i > r_j$ or $r_i = r_j$ and $\nu_i \geqq \nu_j$. We next explore the connection between the domination relation and the production relation.

THEOREM 4.1. *If commodity i produces commodity j then j dominates i.*

Proof. Since i produces j, $(P^m)_{ij} > 0$ for some integer $m \geqq 0$. By the definition of r_j it now follows that for $\alpha > r_j$

$$0 = \lim_{n\to\infty} \alpha^{-n} x(n+m)_j \geqq \left\{ \limsup_{n\to\infty} \alpha^{-n} x(n)_i \right\} (P^m)_{ij} \geqq 0.$$

Since $x(n) \geqq 0$ it follows that $\lim_{n\to\infty} \alpha^{-n} x(n)_i = 0$ for all $\alpha > r_j$, proving that $r_i \leqq r_j$. Next assume that $r_i = r_j \neq 0$, and we will show that $\nu_i \leqq \nu_j$. For every integer $k > \nu_j$,

$$0 = \lim_{n\to\infty} n^{-k} r_j^{-n} x(n+m)_j \geqq \left\{ \limsup_{n\to\infty} n^{-k} r_i^{-n} x(n)_i \right\} (P^m)_{ij} \geqq 0.$$

Since $x(n) \geqq 0$ it follows that $\lim_{n\to\infty} n^{-k} r_i^{-n} x(n)_i = 0$ for all integers $k > \nu_j$, proving that $\nu_i \leqq \nu_j$. If $r_i = r_j = 0$ and $x(n)_i \neq 0$ for some integer $n \geqq 0$, then $x(n+m)_j \geqq x(n)_i (P^m)_{ij} > 0$, showing that $\nu_j \geqq \nu_i + m$, thus completing the proof of Theorem 4.1.

The conclusion of Theorem 4.1 is intuitive. It says that if a commodity which produces some other commodity enjoys a certain growth, then the second commodity enjoys at least that growth.

Theorem 4.1 implies that the domination relation is a class property. An immediate corollary of this fact is the following result:

COROLLARY 4.2. *The geometric growth rate as well as the power growth rate are class properties.*

In the sequel, we will use the relation of domination with respect to classes. We will use the notation r_J (resp., ν_J) for the joint geometric (resp., power) growth rate of the commodities in class J.

Let $(S, x(0), P)$ be a given multiplicative system. We say that a commodity i is *degenerate* if no commodity j with $x(0)_j > 0$ produces i. A class of commodities is called *degenerate* if some, or equivalently every, commodity in the class is degenerate. The set of all degenerate (resp., nondegenerate) commodities will be denoted D (resp., N). The system is called *degenerate* if $D \neq \varnothing$. Of course, a system is degenerate if and only if $x(0)_J = 0$ for some essential class J. A subsystem is called *degenerate* if it is degenerate as an independent system.

We will next obtain a characterization of the growth coefficients.

THEOREM 4.3. *Let $(S, x(0), P)$ be a multiplicative system and J a nondegenerate class of commodities with $r \equiv r_J$. Let $K \equiv N \cap \{i | i \text{ produces } J\}$, $L \equiv N \cap \{i | J \text{ dominates } i\}$ and $M \equiv N \cap \{i | r_i \leqq r\}$. Then $J \subseteq K \subseteq L \subseteq M$ and*

(1)[5] $x(n)_J = x^K(n)_J = x^L(n)_J = x^M(n)_J$.

(2) $r = r(P_K) = r(P_L) = r(P_M)$.

(3) $\nu_J = \nu_r(P_K) = \nu_r(P_L)$.

[5] Since $J \subseteq K \subseteq L \subseteq M$, we identify the commodities of J with the corresponding commodities of the three subsystems.

(4) *ν_J equals the height of J in each of the three subsystems corresponding to K, L and M.*

In addition, if i is a degenerate commodity, then $r_i = \nu_i = x(n)_i = 0$ for all $n = 0, 1, \cdots$.

Proof. We first assume that our system is nondegenerate; i.e., $D = \varnothing$ and $N = \{1, \cdots, S\}$. In this case, obviously, K is an essential set of commodities and by Theorem 4.1 so are L and M. By Lemma 3.1 and the obvious fact that $J \subseteq K \subseteq L \subseteq M$, $x(n)_J = x^K(n)_J = x^L(n)_J = x^M(n)_J$, proving (1). It now follows that the growth coefficients of J in the three subsystems equal the growth coefficients of J in the original system. By (4.6) and Varga (1962, p. 30), $r \leqq r(P_K) \leqq r(P_L) \leqq r(P_M)$. Thus, in order to prove (2) it suffices to show that $r \geqq \rho \equiv r(P_M)$. This result is trivial in the case where $\rho = 0$.

Applying the expansion given in (4.1) for the subsystem corresponding to M, one gets that if $\rho \neq 0$, then

$$\lim_{n\to\infty} \rho^{-n} x^M(n) - \sum_{j=0}^{\nu-1} \binom{n}{j} x^M(0) Q_M^j E_M = 0 \qquad (C, \nu), \tag{4.8}$$

where $\nu \equiv \nu_\rho(P_M)$, $Q_M \equiv \rho^{-1} P_M - I$ and $E_M = E_\rho(P_M)$. By Rothblum (1975, Theorem 3.1, part (3)), there exist a vector $y \in R^{|M|\times 1}$ and an essential class C, with $E_M y = y$ and $(Q_M^{\nu-1} y)_C \gg 0$. By the nondegeneracy assumption, $\alpha \equiv x^M(0) Q_M^{\nu-1} y > 0$. It now follows from (4.8) and arguments similar to those in Hardy (1949, p. 101) that

$$\lim_{n\to\infty} n^{-\nu+1}(\nu-1)\rho^{-n} x^M(n) y = \alpha \qquad (C, \nu).$$

Since $\alpha \neq 0$, it now follows that for some $i \in M$ the limit of $n^{-\nu+1}\rho^{-n} x^M(n)_i$ as $n \to \infty$ is not zero, proving that $r_i \geqq \rho$. By the definition of M, $r \geqq r_i \geqq \rho$, completing the proof of (2) in the nondegenerate case.

Applying a similar analysis to the subsystem corresponding to K shows that if $r \neq 0$, then for some $i \in K$, the limit of $n^{-\nu+1} r^{-n} x^K(n)_i$ as $n \to \infty$ is not zero, where here $\nu \equiv \nu_r(P_K)$. Since $r_i \leqq r(P_K) = r$, the latter implies that $r_i = r$ and $\nu_i \geqq \nu$. If $r = 0$ then by Rothblum (1975, Theorem 3.1, part (3)) there exists a vector $y \in R^{|K|\times 1}$ and an essential class C, such that $(P_K^{\nu-1} y)_C \gg 0$, where (as before) $\nu = \nu_r(P_K)$. By the nondegeneracy assumption, $x^K(\nu-1)y = x^K(0) P_k^{\nu-1} y \neq 0$, which implies that for some $i \in K$, $x(\nu-1)_i \neq 0$. By (4.6), $r_i \leqq r = 0$. Thus, the above implies that in the case where $r = 0$ we also found a state $i \in K$ with $r_i = r$ and $\nu_i \geqq \nu$. By Theorem 4.1 $\nu_J \geqq \nu_i \geqq \nu$, and by (4.7) $\nu_J \leqq \nu$, proving that $\nu_J = \nu$. The proof that $\nu_J = \nu_r(P_L)$ follows similarly, thus completing the proof of (3) in the nondegenerate case.

We next prove (4). By (2), J is a basic class of each of the three subsystems corresponding to K, L and M. By Rothblum (1975, Corollary 3.3), the height of a basic class equals the index of its spectral radius, for the submatrix associated with all commodities producing that class. This implies that the height of J in each of the three subsystems equals $\nu_r(P_K) = \nu(J)$, completing the proof of (4) and Theorem 4.3 for the nondegenerate case.

We next consider the case in which our system is degenerate, i.e., $D \neq \varnothing$. By possibly permuting rows and corresponding columns one can assume that

$$P = \begin{pmatrix} P_D & P_{DB} \\ 0 & P_N \end{pmatrix}.$$

It now follows that for $n = 0, 1, \cdots$

$$P^n = \begin{pmatrix} P_D^n & A_n \\ 0 & P_N^n \end{pmatrix}, \quad \text{where } A_n = \sum_{j=0}^{n-1} P_D^j P_{DB} P_B^{n-j-1},$$

and therefore, since $x(0)_D = 0$,

$$x(n)_D = x(0)_D P_D^n = 0, \tag{4.9}$$

and

$$x(n)_N = x(0)_D A_n + x(0)_N P_N^n = x^N(n). \tag{4.10}$$

It immediately follows from (4.9) and (4.10) that if i is a degenerate commodity then $r_i = \nu_i = 0$ and if J is a nondegenerate class, then r_J and ν_J are the growth rate coefficients of J in the nondegenerate subsystem corresponding to N. The conclusions of (1)–(4) for the degenerate case now follow immediately from the previously proved conclusions for the nondegenerate case. Thus, the proof of Theorem 4.3 is completed.

We pointed out that when a particular commodity is concerned, then the expansion given in (4.1) is not necessarily helpful since we might normalize the n-period output of that commodity by too much. By part (1) of Theorem 4.3, we can apply the expansion (4.1) to a subsystem without changing the n-period output of a particular commodity. Applying our expansion to the subsystem, we next get a more satisfactory expansion of the n-period output.

THEOREM 4.4. *For every class J with $r_J = 0$, $x(n)_J = 0$ for all integers $n \geqq \nu_J$. If J is a class with $r \equiv r_J \neq 0$, put $K \equiv N \cap \{i | i \text{ produces } J\}$. Then*

$$\lim_{n\to\infty} r^{-n} x(n)_K - \sum_{j=0}^{\nu-1} \binom{n}{j} w^j = 0 \qquad (C, \tau) \tag{4.11}$$

and

$$\lim_{m\to\infty} \{r^{-mq+t} x(mq+t)_K - \phi_t(m)\} = 0, \qquad t = 0, \cdots, q-1, \tag{4.12}$$

where $\nu = \nu_J \geqq 1$, $w^j = x(0)_K Q_K^{j-1} E_K$ $(j = 0, \cdots, \nu-1)$, $Q_K = r^{-1} P_k - I$, $E_K \equiv E_r(P_K)$, $\tau = \tau_r(P_K)$, q is a positive integer and $\phi_0, \cdots, \phi_{q-1}$ are vector-polynomials. Moreover, $(w^{\nu-1})_J \gg 0$ and $\max_{0\leqq t\leqq q-1} \deg (\psi_t)_J = \nu - 1$.

Proof. The expansion follows immediately from (4.1) and part (1) of Theorem 4.3. The last conclusion follows from the fact that $\nu_J = \nu$ (part (3) of Theorem 4.3).

Remark. Equation (4.11) can be considered coordinate by coordinate. This enables one to get, for each commodity i, a nonvanishing Cesaro-average polynomial expansion of $r_i^{-n} x(n)_i$ as $n \to \infty$. Now, for each commodity i, let τ_i be the minimal integer τ for which the corresponding expansion holds (C, τ). We mention without proof that τ_i is a class property. Upper bounds on τ_i can be obtained directly from the corresponding results in Rothblum (1980, App. C).

5. Specific results for some structured models. If P is stochastic, then the basic classes are the recurrent classes (e.g., Rothblum (1975)). In the stochastic case, every recurrent class is final, and therefore the height of every basic class is one. Thus, by Theorem 4.3, $\nu_J = 1$ for every recurrent class J and, therefore, $x(n)_J$ converges $(C, 1)$. The $(C, 1)$ limit can be replaced by regular limits when the period of the class is one or ∞.

If we consider the age distribution model, then, in general, there is only one basic class which is the group of ages prior to the end of fertility, and the period of this

class is one. It therefore follows that in (4.1), $\nu = 1$, and that the (C, τ) limit can be replaced by a regular limit.

REFERENCES

D. GALE (1960), *The Theory of Linear Economic Models*, McGraw-Hill, New York.

P. R. HALMOS (1958), *Finite Dimensional Vector Spaces*, Van Nostrand, Princeton, NJ.

G. HARDY (1949), *Divergent Series*, Clarendon Press, Oxford.

T. E. HARRIS (1963), *The Theory of Branching Processes*, Springer-Verlag, Berlin.

R. A. HOWARD AND J. E. MATHESON (1972), *Risk-sensitive Markov decision processes*, Management Sci., 8, pp. 356–269.

S. C. JAQUETTE (1975), *Utility optimal policies in an undiscounted Markov decision process*, Technical Report No. 275, Department of Operations Research, Cornell University, Ithaca, New York.

S. KARLIN (1959), *Mathematical Methods and Theory of Games, Programming and Economics, Vol. I*, Addison-Wesley, Reading, MA.

——— (1966), *A First Course in Stochastic Processes*, Academic Press, New York and London.

J. KEMENY AND J. SNELL (1960), *Finite Markov Chains*, Van Nostrand, Princeton NJ.

L. W. MCKENZIE (1967), *Maximal paths in the von Neumann model*, in Activity Analysis and the Theory of Growth and Planning, E. Malivaud and M.O.L. Bacharach, eds., St. Martin Press, New York.

——— (1971), *Capital accumulation optimal in the final state*, Supplement to Zeitschrift für Nazionale Economie.

C. D. MEYER (1975), *The role of the group inverse in the theory of finite Markov chains*, SIAM Rev., 17, pp. 443–463.

H. NIKAIDO (1968), *Convex Structure and Economic Theory*, Academic Press, New York and London.

U. G. ROTHBLUM (1975), *Algebraic eigenspaces of nonnegative matrices*, Linear Algebra and Appl., 12, pp. 281–292.

——— (1976), *Computation of the eigenprojection of a nonnegative matrix at its spectral radius*, in Mathematical Programming Studies 6, Stochastic Systems: Modeling, Identification and Optimization II, Roger J-B Wets, ed., pp. 188–201.

——— (1981), *Expansions of sums of matrix powers*, SIAM Rev., 23.

——— (1981), *Multiplicative Markov decision chains*, Math. Oper. Res., to appear.

U. G. ROTHBLUM AND A. F. VEINOTT, JR. (1975), *Cumulative average optimality for normalized Markov decision chains*, June, 1975.

J. S. VANDERGRAFT (1968), *Spectral properties of matrices which have invariant cones*, SIAM J. Appl. Math., 16, pp. 1208–1222.

R. S. VARGA (1962), *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 1, March 1981

0196-5212/81/0201-0005 $01.00/0

A GROUP TESTING PROBLEM ON TWO DISJOINT SETS*

GERARD J. CHANG† AND F. K. HWANG‡

Abstract. Recently the following group testing problem has been studied. We have two disjoint sets of items with cardinalities m and n respectively, where each set is known to contain exactly one defective item. The problem is to find the two defective items with a worst-case minimum number of group tests. It was conjectured that $\lceil \log_2 mn \rceil$ tests ($\lceil x \rceil$ denotes the smallest integer not less than x) always suffice. In this paper we prove that the conjecture is true.

1. Introduction. In [1] the following group testing problem was studied. We have two disjoint sets of times $M = \{M_1, \cdots, M_m\}$ and $N = \{N_1, \cdots, N_n\}$, where each set contains exactly one defective item (the others are good items). The problem, called an *(m,n)-problem*, is to find the two defective items by means of a sequence of group tests. A *group test* is a simultaneous test on a group of items with two possible outcomes. The group is identified as *good* if it contains no defective item and identified as *defective* if otherwise. In the latter case the test outcome does not reveal how many or which items are defective.

Let $t_g(m, n)$ denote the number of tests required for the algorithm g to solve the (m, n)-problem, assuming the worst case. Define

$$t(m, n) = \min_g t_g(m, n).$$

It was conjectured in [1] that $t(m, n) = \lceil \log_2 mn \rceil$ (where $\lceil x \rceil$ denotes the smallest integer not less than x), the information-theoretic lower bound, with some partial evidence provided. In this paper we prove the conjecture in its full generality.

2. Some preliminary remarks. A solution of the (m, n)-problem is a pair (M_i, N_j) such that M_i is the defective item in M and N_j the defective item in N. A *solution space* is a set of possible solutions. We also use the notation $(M' \times N')$, $M' \subseteq M$, $N' \subseteq N$, to denote the solution space

$$\{(M_i, N_j) : M_i \in M', N_j \in N'\}.$$

Since the outcome of each test is binary in nature, any algorithm to solve the (m, n)-problem can be represented by a rooted binary tree. At each nonterminal node of the tree, a test is specified and the two links from this node represent the two possible outcomes of the test. The path from the root to any node then indicates a sequence of outcomes for each of the tests made at the nodes along the path (the path on which all outcomes are defective is called the *all-defective path*). Using this information we can associate with each node v a solution space S_v which consists of all possible solutions consistent with the outcomes of the tests on the path from the root to v. For the (m, n)-problem, the solution space associated with the root is $(M \times N)$. Furthermore, the solution space associated with a terminal node is always of cardinality one. From now on we will consider an algorithm always in its binary tree form.

* Received by the editors January 29, 1980, and in revised form July 18, 1980.

† Department of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York.

‡ Bell Laboratories, Murray Hill, New Jersey 07974.

A solution space is said to be *M-distinct* if no two pairs in the space share the same M_i. Let $|X|$ denote the cardinality of the set X. Suppose S is a solution space with

$$|S| = 2^r + 2^{r-1} + \cdots + 2^{r-p} + q,$$

where $2^{r-p-1} > q \geqq 0$. An algorithm g for S is called *M-sharp* if it satisfies the following conditions:

(i) g solves S in $r+1$ tests.

(ii) Let $v(i)$ be the ith node on the all-defective path (the root is labeled as $v(0)$, and let $v'(i)$ be the good son of $v(i)$. Then $|S_{v'(i)}| = 2^{r-i}$ for $i = 0, 1, \cdots, p$.

(iii) $|S_{v(p+1)}| = q$ and $S_{v(p+1)}$ is M-distinct.

If $|S| = 2^r$, then the above conditions are replaced by the single condition

(i') g solves S in r tests.

LEMMA 1. *There exists an M-sharp algorithm for any M-distinct solution space.*

Proof. We can ignore the N-items in a M-distinct solution space. Then it is trivial to find an M-sharp algorithm.

3. The main results. For m fixed, define n_k to be the largest integer such that $mn_k \leqq 2^k$. Then clearly, there exists an n_k whose value is one.

THEOREM 1. *$t(m, n_k) = k$ for all $n_k \geqq 1$. Furthermore, if n_k is odd, then there exists an M-sharp algorithm for the solution space $(m \times n_k)$.*

Proof. By the definition of n_k, $t(m, n_k) \geqq k$. Therefore we need only to show $t(m, n_k) \leqq k$.

The solution space for the $(m, 1)$-problem is M-distinct. Therefore there exists an M-sharp algorithm by Lemma 1. For general $n_k > 1$, we prove Theorem 1 by induction on n_k.

Note that

$$2^{k-2} < mn_{k-1} \leqq 2^{k-1} < m(n_{k-1}+1)$$

implies

$$2^{k-1} < m(2n_{k-1}) \leqq 2^k < m(2n_{k-1}+2).$$

Therefore n_k is either $2n_{k-1}$ or $2n_{k-1}+1$. In the former case we test half of the set N and use induction on the remaining (m, n_{k-1})-problem. In the latter case, let r be the largest integer such that

$$n_k = 2^r n_{k-r} + 1.$$

Then $r \geqq 1$ and n_{k-r} is necessarily odd. Let

$$mn_{k-r} = 2^{k-r-1} + 2^{k-r-2} + \cdots + 2^{k-r-p} + q,$$

where

$$0 \leqq q < 2^{k-r-p-1}.$$

Then

$$\begin{aligned} mn_k &= m(2^r n_{k-r} + 1) \\ &= 2^{k-1} + 2^{k-2} + \cdots + 2^{k-p} + 2^r q + m. \end{aligned}$$

Let g be an M-sharp algorithm for the (m, n_{k-r})-problem. The existence of g is assured by our induction hypothesis. Let v be the node on the all-defective path of g associated with q solutions. Let J be the set of j such that $(M_i, N_j) \in S_v$ for some M_i. For

$j \in J$, let L_j denote a set consisting of those M_i's such that $(M_i, N_j) \in S_v$. Since S_v is M-distinct, the L_j's are disjoint.

We now give an M-sharp algorithm for the (m, n_k)-problem. For easier writing, we will use n for n_{k-r} and n' for n_k. Then N will refer to the set of n items and N_0' to the set of n' items. Partition $N' - \{N_{n'}\}$ into n groups of 2^r items $G_1, \cdots, G_n$. Consider g truncated at the node v; i.e., delete the subtree rooted at v from g. Let g' be an algorithm for the (m, n')-problem where g' is obtained from g by replacing each item N_j in a test by the group G_j and adding $N_{n'}$ to every group tested on the all-defective path. Then each terminal node of g', except the node v' corresponding with v, will be associated with a set of solutions $(M_i \times G_j)$ for some i and j. Since the only uncertainty is on G_j and $|G_j| = 2^r$, r more tests suffice.

Therefore, we need only to give an M-sharp algorithm for the solution space

$$S_{v'} = \left\{ \bigcup_{j \in J} (L_j \times G_j) \right\} \cup (M \times N_{n'}),$$

with $|S_{v'}| = 2^r q + m$. Let $G_{j1}, \cdots, G_{j2^r}$ denote the 2^r items in G_j. Define

$$T_1 = \left\{ \bigcup_{j \in J} G_{j1} \right\} \cup R,$$

where R is a subset of M-items not in any of the L_j, $j \in J$, with $|R| = 2^r q + m - 2^{k-p-1} - q$. Note that there are a total of $m - q$ M-items not in any of the L_j. We now prove that $m - q \geqq |R| > 0$. The former inequality follows immediately from the fact that

$$2^r q < 2^r 2^{k-r-p-1} = 2^{k-p-1}.$$

Furthermore, since

$$\begin{aligned} 2^{k-1} < m(n_{k-1} + 1) &= m(2^{r-1} n_{k-r} + 1) \\ &= 2^{k-2} + 2^{k-3} + \cdots + 2^{k-p-1} + 2^{r-1} q + m, \end{aligned}$$

it follows that

$$2^{r-1} q + m > 2^{k-p-1},$$

or

$$|R| > 2^{r-1} q - q \geqq 0.$$

We test T_1 at $S_{v'}$. Let $S(g)$ and $S(d)$ denote the partition of $S_{v'}$ according as to whether T_1 is good or defective. Then

$$S(d) = \left\{ \bigcup_{j \in J} (L_j \times G_{j1}) \right\} \cup (R \times N_{n'}),$$

with

$$\begin{aligned} |S(d)| &= q + 2^r q + m - 2^{k-p-1} - q \\ &= 2^r q + m - 2^{k-p-1}, \end{aligned}$$

and

$$S(g) = \left\{ \bigcup_{w=2}^{2^r} \bigcup_{j \in J} (L_j \times G_{jw}) \right\} \cup (\{M - R\} \times N_{n'}),$$

with $|S(g)| = |S_{v'}| - |S(d)| = 2^{k-p-1}$. Since $S(d)$ is M-distinct, there exists an M-sharp algorithm for $S(d)$ by Lemma 1. It remains to be shown that $S(g)$ can be done in $k - p - 1$ tests.

Note that $S(g)$ can also be represented as

$$S(g)=\left\{\bigcup_{j\in J}(L_j\times\{G_j-\{G_{j1}\}\cup\{N_{n'}\}\})\right\}\cup\left(\left\{M-R-\bigcup_{j\in J}L_j\right\}\times N_{n'}\right).$$

Since $|G_j-\{G_{j1}\}\cup\{N_{n'}\}|=2^r$, $|M-R-\cup_{j\in J}L_j|$ must also be a multiple of 2^r. Partition $M-R-\{\cup_{j\in J}L_j\}$ into 2^r subsets of equal size, $H_1,\cdots,H_{2^r}$. Define

$$T_w=\left\{\bigcup_{j\in J}G_{jw}\right\}\cup H_w \quad \text{for } w\in W=(2,3,\cdots,2^r\}.$$

Then the T_w's are disjoint. By testing a sequence of proper combinations of T_w's, it is easily seen that in r tests we can partition $S(g)$ into 2^r subsets consisting of

$$S_w=\left\{\bigcup_{j\in j}(L_j\times G_{jw})\right\}\cup(H_w\times N_{n'}),\qquad w=2,\cdots,2^r,$$

and

$$S_1=S(g)-\left\{\bigcup_{w=2}^{2^r}S_w\right\}.$$

Furthermore, S_w is M-distinct and $|S_w|=2^{k-p-r-1}$ for each $w=1,\cdots,2^r$. Therefore each S_w can be solved in $k-p-r-1$ more tests by Lemma 1. This shows that the algorithm just described for $S_{v'}$ is M-sharp. Therefore, the algorithm g' plus the extension on v' as described is M-sharp. The proof of Theorem 1 is complete.

COROLLARY. $t(m, n)=\lceil\log_2 mn\rceil$ *for all* m *and* n.

The corollary follows from Theorem 1 by way of the easily verifiable fact that $t(m, n)$ is monotone nondecreasing in n.

4. Conclusion. In [1], the conjecture on the group testing problem was generalized to a conjecture on bipartite graphs; namely, for every bipartite graph with 2^k edges, there exists a subgraph induced by a subset of vertices with exactly 2^{k-1} edges. It was shown that the truth of the bipartite graph conjecture implies the truth of the group testing conjecture but not the other way around. In this paper we prove the group testing conjecture while the bipartite graph conjecture remains open.

REFERENCE

[1] G. J. CHANG AND F. K. HWANG, *A group testing problem*, SIAM J. Alg. Disc. Meth., 1 (1980), pp. 21–24.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 1, March 1981

0196-5212/81/0201-0006 $01.00/0

THRESHOLD SEQUENCES*

P. L. HAMMER†, T. IBARAKI‡ AND B. SIMEONE†§

Abstract. A graph is threshold if there is a hyperplane separating the characteristic vectors of the independent sets from the characteristic vectors of the nonindependent sets. A sequence of n nonnegative integers is a threshold sequence if it is the degree sequence of a threshold graph with n vertices. Several characterizations of threshold sequences are given, and it is shown that the set of threshold sequences forms a lattice. For an arbitrary degree sequence d (not necessarily threshold), the minimum distance between d and a threshold sequence is called the threshold gap. Its properties are discussed, and the set of threshold sequences at minimum distance from d is also characterized.

1. Introduction. A threshold graph is a graph with the property that there is a hyperplane separating the characteristic vectors of the independent sets from the characteristic vectors of the nonindependent sets of the graph. Threshold graphs, as well as several generalizations and other related concepts, have received wide attention over the last years [1], [3], [4], [6], [7], [8], [9], [10], [11], [15], [16], [17], [19]. They were first introduced in [4], in connection with the equivalence between set packing and knapsack problems, where they were characterized by the property that there are no four vertices a, b, c, d such that (a, b) and (c, d) are edges, while (a, c) and (b, d) are not. Alternatively, they were characterized by the property that the relation defined in the vertex set by "$x \lesssim y$ if and only if all vertices different from y and adjacent to x are also adjacent to y" is a linear pre-order.

The starting point of the present research is the observation that, if d is the degree sequence of a threshold graph (shortly, a *threshold sequence*), then there is only one graph (up to isomorphism) with degree sequence d. It is natural then to ask for a characterization of threshold sequences. Several such characterizations are given in § 3 (the main tool of which is Lemma 9) using the basic properties of degree sequences prepared in § 2. One of them states that threshold sequences are precisely those which satisfy the classical Erdös-Gallai inequalities as equalities. Hence threshold sequences are among the graphic sequences the "least" graphic ones in a well-defined sense.

In a previous paper, it has been observed that a graph is split if and only if its degree sequence satisfies the last Erdös-Gallai inequality as an equality [11]. Hence, our result gives an algebraic explanation of the known fact [4], [8], that a threshold graph is split.

Various characterizations of the threshold sequences, given in § 3, naturally lead to the introduction of a parameter, the threshold gap of a sequence, with the property that a sequence is threshold if and only if its threshold gap is zero. In § 5, we show that the threshold gap of a sequence is actually the minimum distance (in a suitable norm) between d and any threshold sequence of the same length. This result substantiates the interpretation of the threshold gap as a measure of "non-thresholdness" of a sequence. As discussed in § 4, the set of all threshold sequences of length n is seen to have a lattice structure with respect to the operations consisting in taking the minimum and, respec-

* Received by the editors January 2, 1980, and in revised form June 3, 1980. The research was supported by the National Research Council of Canada under Grant A8552, a Canada Council Grant for Exchange of Scientists between Japan and Canada, and the Italian National Research Council. A preliminary version of this note was presented at the Ninth Southeastern Conference on Combinatorics, Graph Theory and Computing, 1978.

† Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

‡ Department of Applied Mathematics and Physics, Kyoto University, Kyoto, Japan.

§ Istituto per le Applicazioni del Calcolo, C.N.R., Viale del Policlinico 137, 00161, Rome, Italy.

tively, the maximum componentwise of two sequences. This property plays a role in the characterization of the set of threshold sequences at minimum distance from a given arbitrary sequence d. Among such closest threshold sequences, there are two special ones which are obtained from d by replacing its "head" or its "feet" so as to transform d into a threshold sequence. Properties of these sequences are discussed in some detail.

2. Some properties of integer sequences. In preparation to the study of threshold sequences, some preliminary results on integer sequences are needed.

Let $d = (d_1, \cdots, d_n)$ be a sequence of integers such that $n-1 \geqq d_1 \geqq \cdots \geqq d_n \geqq 0$. Such a sequence will be called *proper*. A proper sequence is *graphic* if there is some graph with degree sequence d. Necessary and sufficient conditions for d to be graphic are [5]:

(a) $\sum_{i=1}^{n} d_i$ is even;

(b) For $k = 1, \cdots, n-1$,

$$\sum_{i=1}^{k} d_i \leqq k(k-1) + \sum_{i=k+1}^{n} \min\{k, d_i\}. \tag{1}$$

The inequality (1) will be called the kth *Erdös-Gallai inequality* (EGI). Let us define

$$m = m(d) = \max\{k : d_k \geqq k-1\}. \tag{2}$$

PROPOSITION 1. *One has $d_k \geqq m-1$ for $k \leqq m$ and $d_k \leqq m-1$ for $k > m$. Therefore*

$$\sum_{i=k+1}^{n} \min\{k, d_i\} = \sum_{i=k+1}^{n} d_i \quad \text{for } k \geqq m. \tag{3}$$

THEOREM 2 (Li [14]). *If the mth EGI holds, then the kth EGI is automatically satisfied for $m-1 \leqq k \leqq n$.*

From now on, we shall deal only with the reduced system formed by the first m Erdös-Gallai inequalities, the remaining ones being redundant.

Another characterization of a graphic sequence is given by the next proposition. The proof is straightforward and can be found in [12].

PROPOSITION 3. *A proper sequence d with even sum is graphic if and only if*

$$\sum_{j=1}^{h} \max\{h-1, d_j\} \leqq h(h-1) + \sum_{j=h+1}^{n} d_j \quad \text{for all } h \text{ with } m \leqq h \leqq n-1. \tag{4}^{1}$$

If d is a proper sequence, let us define, for $k = 1, \cdots, n-1$,

$$d_k = |\{i : d_i \geqq k\}| = \max\{i : d_i \geqq k\}. \tag{5}$$

It is well known (see e.g. [2, Chap. 6]) that d_k^* can be interpreted as the number of points in the kth row of the *Ferrer diagram* of d. As an example, Fig. 1 shows the Ferrer diagram of the sequence $d = (5, 5, 3, 2, 2, 1)$. Here $d^* = (6, 5, 3, 2, 2)$.

We remark here that many properties of degree sequences have simple geometrical interpretations in terms of Ferrer diagrams. For example, one has $\sum_{i=1}^{n} d_i = \sum_{k=1}^{n-1} d_k^*$ (e.g., see [2]) for any proper sequence d. Actually both $\sum_{i=1}^{n} d_i$ and $\sum_{k=1}^{n-1} d_k^*$ are equal to the total number of points in the Ferrer diagram of d.

[1] We make the convention that, when $h > k$, $\sum_{i=h}^{k} a_i = 0$ and $[h, k]$ is the empty interval.

```
2 . .
2 . .
3 . . .
5 . . . . .
6 . . . . . .
  5 5 3 2 2 1
```

FIG. 1. *Example of Ferrer diagram.*

It is easy to see (taking into account Proposition 1) that

$$d_k^* = m + |\{i: m+1 \leqq i \leqq n, d_i \geqq k\}| \quad \text{for } 1 \leqq k \leqq m-1, \tag{6}$$

and conversely,

$$\begin{aligned} &d_k = m-1 && \text{for } m < k \leqq d_m^* - 1 \\ &d_k = m-2 && \text{for } d_{m-1}^* < k \leqq d_{m-2}^*, \\ &\quad\vdots \\ &d_k = 1 && \text{for } d_2^* < k \leqq d_1^*, \\ &d_k = 0 && \text{for } d_1^* < k \leqq n. \end{aligned} \tag{7}$$

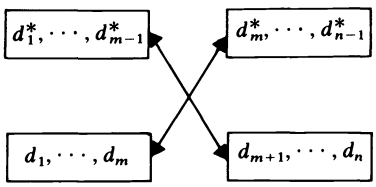
Similarly, for $1 \leqq k \leqq m$,

$$d_k = m - 1 + |\{i: m \leqq i < n, d_i^* \geqq k\}|, \tag{8}$$

and conversely

$$\begin{aligned} &d_k^* = m && \text{for } m-1 < k \leqq d_m, \\ &d_k^* = m-1 && \text{for } d_m < k \leqq d_{m-1}, \\ &\quad\vdots \\ &d_k^* = 1 && \text{for } d_2 < k \leqq d_1, \\ &d_k^* = 0 && \text{for } d_1 < k \leqq n-1. \end{aligned} \tag{9}$$

Relations (6), (7), (8) and (9) show that, once m is known, the vector $(d_1^*, \cdots, d_{m-1}^*)$ depends only on the vector $(d_{m+1}, \cdots, d_n)$ and vice versa. Likewise, the vector $(d_1, \cdots, d_m)$ depends only on the vector $(d_m^*, \cdots, d_{n-1}^*)$ and vice versa. The situation is schematized in the following diagram,



Let us put

$$\phi_k = \phi_k(d) = \begin{cases} k(k-1) + \sum_{i=k+1}^{n} \min\{k, d_i\} - \sum_{i=1}^{k} d_i & \text{for } 1 \leqq k \leqq m, \\ k(k-1) + \sum_{i=k+1}^{n} d_i - \sum_{i=1}^{k} \max\{k-1, d_i\} & \text{for } m \leqq k \leqq n. \end{cases}$$

(Notice that, for $k = m$, there is no conflict between the two different definitions of ϕ_m). For $1 \leqq k \leqq m$, ϕ_k is just the slack of (1); for $m \leqq h \leqq n-1$, ϕ_h is the slack of (4).

PROPOSITION 4. *One has*

$$\begin{aligned} &\phi_1 = \Delta_1, \\ &\phi_2 = \phi_1 + \Delta_2, \\ &\cdots \\ &\phi_n = \phi_{n-1} + \Delta_n, \end{aligned} \tag{10}$$

where

$$\Delta_k = \Delta_k(d) = \begin{cases} d_k^* - d_k - 1 & \text{for } 1 \leqq k \leqq m-1, \\ m - 1 - d_m & \text{for } k = m, \\ d_{k-1}^* - d_k & \text{for } m+1 \leqq k \leqq n. \end{cases} \tag{11}$$

Proof. One has $\phi_1 = \sum_{i=2}^{n} \min\{d_i, 1\} - d_1 = d_1^* - d_1 - 1 = \Delta_1$. Noticing that, for any $1 \leqq k \leqq m-1$,

$$\sum_{i=k+1}^{n} \min\{k, d_i\} = \sum_{i=k+1}^{d_k^*} \min\{k, d_i\} + \sum_{i=d_k^*+1}^{n} \min\{k, d_i\}$$

$$= (d_k^* - k)k + \sum_{i=d_k^*+1}^{n} d_i,$$

one has, for $k = 2, \cdots, m-1$,

$$\begin{aligned} \phi_k - \phi_{k-1} &= k(k-1) - (k-1)(k-2) + (d_k^* - k)k - (d_{k-1}^* - k + 1)(k-1) \\ &\quad + \sum_{i=d_k^*+1}^{n} d_i - \sum_{i=d_{k-1}^*}^{n} d_i - \sum_{i=1}^{k} d_i + \sum_{i=1}^{k-1} d_i \\ &= (d_k^* - d_{k-1}^*)k + d_{k-1}^* + \sum_{i=d_k^*+1}^{d_{k-1}^*} d_i - d_k. \end{aligned}$$

Since $d_i = k-1$ for $d_k^* + 1 \leqq i \leqq d_{k-1}^*$, one has, for $k = 2, \cdots, m-1$,

$$\phi_k - \phi_{k-1} = (d_k^* - d_{k-1}^*)k + d_{k-1}^* - 1 - (d_k^* - d_{k-1}^*)(k-1) - d_k = d_k^* - d_k - 1.$$

On the other hand,

$$\begin{aligned} \phi_m - \phi_{m-1} &= \left\{ m(m-1) + \sum_{i=m+1}^{n} d_i - \sum_{i=1}^{m} d_i \right\} - \left\{ (m-1)^2 + \sum_{i=m+1}^{n} d_i - \sum_{i=1}^{m-1} d_i \right\} \\ &= m - 1 - d_m. \end{aligned}$$

Moreover, noticing that, for any k such that $m+1 \leqq k \leqq n-1$, one has

$$\sum_{i=1}^{k} \max\{k-1, d_i\} = \sum_{i=1}^{d_{k-1}^*} \max\{k-1, d_i\} + \sum_{i=d_{k-1}^*+1}^{k} \max\{k-1, d_i\}$$

$$= \sum_{i=1}^{d_{k-1}^*} d_i + (k - d_{k-1}^*)(k-1),$$

one must also have, for $m \leqq k \leqq n$,

$$\begin{aligned}\phi_{k+1}-\phi_k &= (k+1)k-k(k-1)+\sum_{j=k+2}^{n} d_j-\sum_{j=k+1}^{n} d_j-\sum_{j=1}^{d^*_{k+1}} d_j+\sum_{j=1}^{d^*_k} d_j \\ &\quad -(k+1-d_{k+1})h+(k-d^*_k)(k-1) \\ &= -d_{k+1}+\sum_{j=d^*_{k+1}+1}^{d^*_k} d_j+k(d^*_{k+1}-d_k)+d^*_k \\ &= -d_{k+1}+(d^*_k-d^*_{k+1})k+k(d^*_{k+1}-d^*_k)+d^*_k \\ &= d^*_k-d_{k+1}. \qquad \square\end{aligned}$$

The following two propositions exhibit important properties of Δ_k defined in (11). Proofs are omitted for simplicity. See [12] for details.

PROPOSITION 5. *If d is any proper sequence, one has*

$$\sum_{k=1}^{n} \Delta_k = 0, \tag{12}$$

and

$$\sum_{k=1}^{m} |\Delta_k| = \sum_{k=m+1}^{n} |\Delta_k|. \tag{13}$$

PROPOSITION 6. *If d is any proper sequence such that $\Delta_i \equiv d^*_i - d_i - 1 = 0$ for $i = 1, \cdots, m-1$, then d is graphic if and only if $d_m = m-1$.*

3. Characterizations of the threshold sequences. For a graph $G=(V, E)$, the degree of vertex v in G is denoted by $d(v) = d_G(v)$. It is assumed throughout this section that the vertices of G are numbered from 1 to n in such a way that if d_i is the degree of vertex i one has $d_i \geqq d_2 \geqq \cdots \geqq d_n$.

If v is a vertex of G, the *neighborhood* of v is the set $N(v)$ of all vertices adjacent to v. Following [8], we introduce in V a linear preorder $\succcurlyeq$ by: $u \succcurlyeq v \Leftrightarrow N(u)-\{v\} \supseteqq N(v)-\{u\}$. It is easy to see that

LEMMA 7. *If G is threshold, $u \preccurlyeq v$ if and only if $d(u) \leqq d(v)$.*

Following the terminology in [8], we call a *critical nonthreshold* (CNT) *configuration* of G any set of four vertices u, v, w, z such that $(u, v) \in E$, $(w, z) \in E$, $(u, w) \notin E$, $(v, z) \notin E$. Given such a CNT configuration, the operation of removing edges (u, v) and (w, z) and adding the edges (u, w) and (v, z) is called an *interchange*.

LEMMA 8. *If d is a threshold sequence, there is a unique (up to isomorphism) graph with degree sequence d.*

Proof. By a well-known theorem of Ryser [18, p. 68], any two graphs with the same degree sequence can be obtained each from the other through a finite sequence of interchanges. If G is a threshold graph with degree sequence d, no interchange can be performed, because CNT configurations are forbidden in a threshold graph [4]. Hence G is the only graph (up to isomorphism) with degree sequence d. $\square$

For every pair (h, k) of positive integers, let N^h_k denote the set of the first h integers starting from 1 and excluding k; i.e.,

$$N^h_k = \begin{cases} \{1, 2, \cdots, h\} & \text{if } h < k, \\ \{1, 2, \cdots, k-1, k+1, \cdots, h+1\} & \text{if } h \geqq k. \end{cases} \tag{14}$$

We define N^0_k to be the empty set.

The following characterization of adjacency in threshold graphs is very useful in deriving some of the main results below.

LEMMA 9. *A graph G with degree sequence d is threshold if and only if*

$$N(k) = N_k^{d_k} \quad \text{for } k = 1, 2, \cdots, n. \tag{15}$$

Proof. Let G be threshold. Since $|N(k)| = d_k$ for $1 \leqq k \leqq n$, (15) is proved if $(j, k) \in E$ implies $(i, k) \in E$ for all $i < j$ and $i \neq k$. Indeed, for $i < j$, one has $d_i \geqq d_j$; hence $i \succcurlyeq j$ by Lemma 7, so that $(j, k) \in E$ implies $(i, k) \in E$ for $i \neq k$.

Conversely, $i < j \Rightarrow d_i \geqq d_j \Rightarrow N_i^{d_i} - \{i\} \supseteq N_j^{d_j} - \{i\} \Rightarrow i \succcurlyeq j$, and hence $\succcurlyeq$ is a linear pre-order, i.e., G is threshold. □

LEMMA 10. *A graph G with degree sequence d is threshold if and only if*

$$\Delta_k = d_k^* - d_k - 1 = 0 \quad \text{for } k = 1, 2, \cdots, m-1. \tag{16}$$

Proof. Let G be threshold and $1 \leqq k \leqq m-1$. For $i > k$, notice that $d_i \geqq k \Rightarrow \{1, 2, \cdots, k\} \subseteq N_i^{d_i}$ (by (14)) $\Rightarrow k \in N(i)$ (by Lemma 9) $\Rightarrow i \in N(k) \Rightarrow i \in N_k^{d_k}$ (by Lemma 9). Therefore

$$\begin{aligned} d_k^* &= \max\{i: d_i \geqq k\} \qquad \text{(by definition (5))} \\ &= \max\{i: i > k, d_i \geqq k\} \qquad \text{(since } d_m \geqq m-1 \geqq k\text{)} \\ &= \max\{i: i \in N_k^{d_k}\} = d_k + 1. \end{aligned}$$

Conversely, assume that (16) holds and define a graph $H = (V, F)$, where $V = \{1, 2, \cdots, n\}$ and

$$F = \{(i, j): 1 \leqq i \leqq j \leqq m\} \cup \{(i, j): m+1 \leqq j \leqq n, 1 \leqq i \leqq d_j\}. \tag{17}$$

We claim that $d_H(k) = d_k$ for $k = 1, 2, \cdots, n$. This is obvious when $m+1 \leqq k \leqq n$. Assume then that $1 \leqq k \leqq m$. By construction, one has $d_H(k) = m - 1 + |\{j: m+1 \leqq j \leqq n, k \leqq d_j\}|$. In particular, $d_H(m) = m-1$ for $k = m$. On the other hand, since d is graphic and satisfies (16), one has $d_m = m-1$ (i.e., $\Delta_m = 0$) by Proposition 6. Hence $d_H(m) = d_m$. For $1 \leqq k \leqq m-1$, one has $d_H(k) = d_k^* - 1$ from (6). Hence $d_H(k) = d_k$ for $k = 1, 2, \cdots, m-1$ by (16).

Next we observe that by construction $N_H(k) = N_k^{d_k}$ for all k; hence H is threshold by Lemma 9. Since H and G have the same degree sequence, they are isomorphic by Lemma 8. Hence G is threshold. □

THEOREM 11. *A proper sequence is threshold if and only if one of the following conditions is satisfied*:

(i) $\Delta_i = 0$ *for* $i = 1, 2, \cdots, m$.
(ii) $\phi_i = 0$ *for* $i = 1, 2, \cdots, m$.
(iii) $\Delta_i = 0$ *for* $i = m+1, m+2, \cdots, n$.
(iv) $\phi_i = 0$ *for* $i = m+1, m+2, \ldots, n$.

Proof. If d is threshold, $\Delta_i = 0$ for $i = 1, 2, \cdots, m-1$ by Lemma 10. $\Delta_m = 0$ also holds by Proposition 6 since d is graphic. Conversely, if $\Delta_i = 0$ for $i = 1, 2, \cdots, m$, d is graphic by Proposition 6, and threshold by Lemma 10. This proves (i). (ii) is obviously equivalent to (i) by Proposition 4. The equivalence between (i) and (iii) follows from Proposition 5. (iii) and (iv) are equivalent by Proposition 4. □

Theorem 11 implies that threshold sequences are, among the graphic sequences, "the least" graphic ones in a well-defined sense. The same result could have been obtained by combining together Li [14, Theorems 18 and 19]. However, our approach is entirely different from Li's, which is based on the concept of uniquely realizable degree sequences.

4. The lattice of threshold sequences. For any two proper sequences d and d', we define two sequences $d \vee d'$ and $d \wedge d'$, which are also proper, by

$$(18)\qquad \begin{aligned}(d \vee d')_i &= \max\{d_i, d'_i\}, \qquad i = 1, 2, \cdots, n,\\ (d \wedge d')_i &= \min\{d_i, d'_i\}, \qquad i = 1, 2, \cdots, n.\end{aligned}$$

The set of all threshold sequences of length n will be denoted by $\mathcal{T}_n$.

LEMMA 12. *If $c, c' \in \mathcal{T}_n$, then $c \vee c'$, $c \wedge c' \in \mathcal{T}_n$.*

Proof. Let the threshold graphs $G = (V, E)$ and $G' = (V, E')$ have degree sequences c and c' respectively. Then

$$N_G(k) = N_k^c \quad \text{and} \quad N_{G'}(k) = N_k^{c'}, \qquad k = 1, 2, \cdots, n$$

by Lemma 9. Let $c'' = c \vee c'$ and $G'' = (V, E \cup E')$. Then

$$\begin{aligned}N_{G''}(k) &= N_G(k) \cup N_{G'}(k)\\ &= N_k^{\max\{c_k, c'_k\}} = N_k^{c''}, \qquad k = 1, 2, \cdots, n.\end{aligned}$$

Then G'' is a threshold graph by Lemma 9 and has degree sequence c''. Thus $c'' \in \mathcal{T}_n$. The other half of the lemma, i.e., $c \wedge c' \in \mathcal{T}_n$, can be similarly proved by considering

The next result is an immediate consequence of this lemma.

THEOREM 13. *$(\mathcal{T}_n, \vee, \wedge)$ is a finite lattice with maximum element $(n-1, n-1, \cdots, n-1)$ and minimum element $(0, 0, \cdots, 0)$.*

5. Threshold sequences at minimum distance from a graphic sequence. Define the *threshold gap* of a proper sequence d by

$$(19)\qquad t(d) = \frac{1}{2}\sum_{i=1}^{m} |\Delta_i| = \frac{1}{2}\sum_{i=m+1}^{n} |\Delta_i| \qquad \text{(see Proposition 5).}$$

$t(d)$ is a measure of "nonthresholdness" of d. Indeed, by Theorem 11 and Proposition 5, a *proper sequence d is threshold if and only if $t(d) = 0$*. The aim of the present section is to make this statement more precise.

For any two proper sequences d and d', let us put

$$(20)\qquad \|d - d'\| = \frac{1}{2}\sum_{i=1}^{n} |d_i - d'_i|.$$

We notice that $\|d - d'\|$ is a nonnegative integer if d and d' are graphic sequences. Indeed, if $I = \{i\colon d_i \geqq d'_i\}$ and $\bar{I} = \{1, 2, \cdots, n\} - I$, one has

$$\sum_{i=1}^{n} |d_i - d'_i| = \left(\sum_{i\in I} d_i - \sum_{i\in I} d'_i\right) + \left(\sum_{i\in \bar{I}} d'_i - \sum_{i\in \bar{I}} d_i\right).$$

Since $\sum_{i=1}^{n} d_i$ is even, $\sum_{i\in I} d_i$ and $\sum_{i\in I} d_i$ have the same parity. The same is true for $\sum_{i\in I} d'_i$ and $\sum_{i\in I} d'_i$; hence $\sum_{i=1}^{n} |d_i - d'_i|$ is even.

The main result of this section (Theorem 18) shows that the threshold gap $t(d)$ is the minimum distance $\|\cdot\|$ between d and any threshold sequence c.

To start with, following a procedure pioneered by Procrustes of Eleusis, we associate to a proper sequence d two canonical sequences $\hat{d}$ and $\check{d}$ obtained by replacing

the "head" or the "feet" of d, respectively, in the following way:

$$\begin{aligned} &\hat{d}_k = d_k + \Delta_k, && k = 1, 2, \cdots, m, \\ &\hat{d}_k = d_k, && k = m+1, m+2, \cdots, n, \\ &\check{d}_k = d_k, && k = 1, 2, \cdots, m, \\ &\check{d}_k = d_k + \Delta_k, && k = m+1, m+2, \cdots, n. \end{aligned} \tag{21}$$

From the definition of Δ_k, one has

$$\begin{aligned} &\hat{d}_k = d_k^* - 1 \quad \text{for } 1 \leqq k \leqq m-1, \\ &\hat{d}_m = m-1, \\ &\check{d}_m = d_m, \\ &\check{d}_k = d_{k-1}^* \quad \text{for } m+1 \leqq k \leqq n. \end{aligned} \tag{22}$$

Since $n-1 \geqq d_1^* - 1 \geqq d_2^* - 1 \geqq \cdots \geqq d_{m-1}^* - 1 \geqq m-1 \geqq d_{m+1} \geqq \cdots \geqq d_n \geqq 0$, and $n-1 \geqq d_1 \geqq \cdots \geqq d_m \geqq d_m^* \geqq \cdots \geqq d_{n-1}^* \geqq 0$ from (6)–(9), $\check{d}$ and $\check{d}$ are also proper sequences. One has

$$\begin{aligned} &m(\hat{d}) = m, \\ &m(\check{d}) = M = \begin{cases} m & \text{if } d_m = m-1, \\ m+1 & \text{if } d_m > m-1. \end{cases} \end{aligned} \tag{23}$$

The first relation is obvious. The second relation follows from the observation that

$$\begin{aligned} &d_m = m-1 \Rightarrow d_m = m-1 \geqq d_m^* = \check{d}_{m+1}, \\ &d_m \geqq m \Rightarrow \check{d}_{m+1} = d_m^* \geqq m \geqq d_{m+1}^* = \check{d}_{m+2}. \end{aligned}$$

LEMMA 14. *For a proper sequence d, $\hat{d}$ and $\check{d}$ are both threshold sequences.*

Proof. First of all, we notice that $\hat{d}$ depends only on $(d_{m+1}, \cdots, d_n)$ because of (6) and (9). Also $\hat{d}_k^* = d_k^*$ for $k = 1, 2, \cdots, m-1$ by (21). Since $\Delta_k(\hat{d}) = \hat{d}_k^* - \hat{d}_k - 1 = d_k^* - d_k^* + 1 - 1 = 0$ for $k = 1, 2, \cdots, m-1$, and $\Delta_m(\hat{d}) = m - 1 - \hat{d}_m = 0$ by (22), $\hat{d}$ is threshold by Theorem 11(i). Similarly for $\check{d}$. □

LEMMA 15. *For a proper sequence d, $\hat{d}$ and $\check{d}$ satisfy*

$$\|d - \hat{d}\| = \|d - \check{d}\| = t(d). \tag{24}$$

Proof. By definitions (19) and (21),

$$\|d - \hat{d}\| = \frac{1}{2} \sum_{k=1}^{m} |\Delta_k| = t(d).$$

Similarly for $\check{d}$. □

This lemma provides an alternative proof of the fact that $t(d)$ is a nonnegative integer if d is a graphic sequence.

We need two more lemmas before presenting the main result of this section. The first one can be easily proved by using Lemma 9; hence the proof is omitted.

LEMMA 16. *Let $G = (V, E)$ be a threshold graph. Let h, k be vertices of V, with $h < k$. Then the following graphs G_1, G_2 and G_3 are also threshold:*

$$G_1 = (V, E - \{(h, k)\}) \quad \text{if } h = \max\{i : (i, k) \in E\} \text{ or } k = \max\{i : (h, i) \in E\},$$

$$G_2 = (V, E \cup \{(h+1, k)\}) \quad \text{if } h = \max\{i : (i, k) \in E\} \text{ and } h + 1 < k,$$

$$G_3 = (V, E \cup \{(h, k+1)\}) \quad \text{if } k = \max\{i : (h, i) \in E\} \text{ and } k < n.$$

We remark that $h = \max\{i: (i, k) \in E\}$ holds if $h = d_k < k$, and $k = \max\{i: (h, i) \in E\}$ holds if $k = d_h + 1 > h$. In some cases, the vertices of the resulting graph may have to be renumbered in order to obtain a proper sequence, since the deletion or addition of an edge changes its degree sequence.

LEMMA 17. *Let c, $c' \in \mathcal{T}_n$ and $m = m(c) = m(c')$. Then $c = c'$ if $c_i = c'_i$ for $i = 1, 2, \cdots, m$, or $i = m+1, m+2, \cdots, n$.*

Proof. Assume that $c_i = c'_i$ for $i = m+1, \cdots, n$; the other case can be similarly treated.

Since $\Delta_i(c) = c_i^* - c_i - 1 = 0$ and $\Delta_i(c') = (c'_i)^* - c'_i - 1 = 0$ for $i = 1, 2, \cdots, m-1$ by Theorem 11(i), and noticing that c_i^* and $(c'_i)^*$ for $i = 1, 2, \cdots, m-1$ are determined by $(c_{m+1}, \cdots, c_n) = (c'_{m+1}, \cdots, c'_n)$, one has $c_i = c'_i$ for $i = 1, 2, \cdots, m-1$. Moreover, $c_m = c'_m = m-1$ because $\Delta_m(c) = \Delta_m(c') = 0$. Thus $c = c'$. □

THEOREM 18. *One has*

$$t(d) = \min_{c \in \mathcal{T}_n} \|d - c\|. \tag{25}$$

Proof. Let $\tilde{c} \in \mathcal{T}_n$ be optimal, i.e., $\|d - \tilde{c}\| = \min_{c \in \mathcal{T}_n} \|d - c\|$. We first show that $m \leqq \tilde{m} \leqq M$, where $m = m(d)$, $\tilde{m} = m(\tilde{c})$ and M is defined by (23). Assume that $\tilde{m} < m$. Setting $h = \tilde{c}_{\tilde{m}+1} \leqq \tilde{m} - 1$, add a new edge $(h+1, \tilde{m}+1)$ to the threshold graph $\tilde{G}$ having degree sequence $\tilde{c}$. The new graph G' so obtained is also threshold by Lemma 16, and has degree sequence c' with $c'_i = \tilde{c}_i + 1$ for $i = h+1, \tilde{m}+1$ and $c'_i = c_i$ for $i \neq h+1, \tilde{m}+1$. By the definition of h, one has $\tilde{c}_h > \tilde{c}_{h+1}$. In addition assume that $\tilde{c}_{\tilde{m}} \geqq \tilde{m} - 1 > \tilde{c}_{\tilde{m}+1}$. Then the addition of edge $(h+1, \tilde{m}+1)$ does not require any renumbering of vertices in G' (see the remark after Lemma 16). Since $c'_{\tilde{m}+1} = \tilde{c}_{\tilde{m}+1} + 1 \leqq \tilde{m} \leqq m - 1 \leqq d_m \leqq d_{\tilde{m}+1}$, one has $|c'_{\tilde{m}+1} - d_{\tilde{m}+1}| = |c_{\tilde{m}+1} - d_{\tilde{m}+1}| - 1$, while certainly $|c'_{h+1} - d_{h+1}| \leqq |\tilde{c}_{h+1} - d_{h+1}| + 1$. Hence $\|c' - d\| \leqq \|\tilde{c} - d\|$. Repeating the same procedure (if necessary) with $\tilde{c}$ replaced by c', one eventually obtains $\tilde{c}_{\tilde{m}+1} = \tilde{m} - 1$. In this case, $c'_{h+1} = c'_{\tilde{m}+1} = \tilde{m}$ holds since $h + 1 = \tilde{m}$ and $c'_{\tilde{m}} = \tilde{c}_{\tilde{m}+1} = \tilde{m}$ (recall that $\tilde{c}$ is threshold). Thus the renumbering of vertices of G' is not necessary. $|c'_{\tilde{m}+1} - d_{\tilde{m}+1}| = |\tilde{c}_{\tilde{m}+1} - d_{\tilde{m}+1}| - 1$ holds as before, and one has $|c'_{\tilde{m}} - d_{\tilde{m}}| = |\tilde{c}_{\tilde{m}} - d_{\tilde{m}}| - 1$ by $c'_{\tilde{m}} = \tilde{m} \leqq m - 1 \leqq d_m \leqq d_{\tilde{m}}$. Therefore $\|c' - d\| < \|\tilde{c} - d\|$ and $\tilde{c}$ is not optimal, a contradiction. A similar argument applies when $\tilde{m} = m(\tilde{c}) > m$; we can show that $\tilde{c}$ is not optimal.

Now assume that $m \leqq \tilde{m} \leqq M$. We shall show that $\|\tilde{c} - d\| \geqq \|\hat{d} - d\|$ if $\tilde{m} = m$, and $\|\tilde{c} - d\| \geqq \|\check{d} - d\|$ if $\tilde{m} = M$. We consider only the case $\tilde{m} = m$; the other case can be dealt with in a similar way. If $\tilde{c} \neq \hat{d}$, then $\tilde{c}_i \neq \hat{d}_i$ for some $m+1 \leqq i \leqq n$ by Lemma 17. Let k be the maximum index i (if any) such that $m+1 \leqq i \leqq n$ and $\tilde{c}_i > d_i$. For $h = \tilde{c}_k (< k)$, remove edge (h, k) from $\tilde{G}$ to obtain a threshold graph G' (by Lemma 16) with degree sequence c', where $c'_i = \tilde{c}_i - 1$ for $i = h, k$ and $c'_i = \tilde{c}_i$ for $i \neq h, k$. The renumbering of indices is not necessary in this case, as is easily shown. Since $c'_k = \tilde{c}_k - 1 \geqq \hat{d}_k = d_k$, one has $|c'_k - d_k| = |\tilde{c}_k - d_k| - 1$; hence $\|c' - d\| \leqq \|\tilde{c} - d\|$ because in any case $|c'_h - d_h| \leqq |\tilde{c}_h - d_h| + 1$. $m(c') = m$ is also obvious. By repeated applications of the above procedure, one eventually has a threshold sequence c' such that $m(c') = m$, $c'_i \leqq \hat{d}_i$ for $i = m+1, \cdots, n$, and $\|c' - d\| \leqq \|\tilde{c} - d\|$.

If one still has $c' \neq \hat{d}$, there must be some i, $m+1 \leqq i \leqq n$, such that $c'_i < \hat{d}_i$; let k be the maximum such i. The addition of edge $(h+1, k)$, where $h = c'_k (\leq m - 1 < k - 1)$, results in a new threshold graph G'' (by Lemma 16) with degree sequence c'', where $c''_i = c'_i + 1$ for $i = h+1, k$ and $c''_i = c'_i$ for $i \neq h+1, k$. Since $c''_k = c'_k + 1 \leqq \hat{d}_k = d_k$, one has $\|c'' - d\| \leqq \|c' - d\|$ as before.

In conclusion, by repeated application of the above procedures, one eventually gets a final threshold sequence c such that $\|c - d\| \leqq \|\tilde{c} - d\|$ and $c_i = \hat{d}_i = d_i$ for $i =$

$m+1, \cdots, n$. Then $c = \hat{d}$ by Lemma 17, showing that $\|\hat{d} - d\| \leqq \|c - d\|$ for all $c \in \mathcal{T}_n$. But then $t(d) = \|\hat{d} - d\| = \min_{c \in \mathcal{T}_n} \|c - d\|$ by Lemma 15. □

Given a proper sequence d, denote by $\mathcal{T}_n(d)$ the set of threshold sequences c of length n satisfying $\|c - d\| = t(d)$, i.e., having minimum distance from d. Furthermore, let

$$d^+ = \hat{d} \vee \check{d}, \; d^- = \hat{d} \wedge \check{d}. \tag{26}$$

Both d^+ and d^- are threshold sequences by Lemmas 14 and 12, and they satisfy

$$\|d^+ - d\| = \|d^- - d\| = t(d), \tag{27}$$

i.e., $d^+, d^- \in \mathcal{T}_n(d)$, as easily proved by (21) and Proposition 5.

LEMMA 19. *Let d be a proper sequence, and let d^+ and d^- be defined as above. Then*

$$d_k^- < d_k \Rightarrow d_h^+ > d_h \quad \textit{for every } h \textit{ such that } d_k^- < h \leqq d_k^+,$$

$$d_k^+ > d_k \Rightarrow d_h^- < d_h \quad \textit{for every } h \textit{ such that } d_k^- < h \leqq d_k^+.$$

Proof. We shall prove the lemma only for the case $d_k^- < d_k$ and $m+1 \leqq k < n$. The proofs for the other cases are similar. Notice first that $d_k^- < d_k$ implies $d_k^- = \check{d}_k = d_{k-1}$ and $d_k^+ = \check{d}_k = d_k$. If $d_k^- < h \leqq d_k^+$, then $d_{k-1}^* < h \leqq d_k \leqq m-1$. One has $d_h^* \geqq k$ by $d_k \geqq h$ and (5), and $d_h < k-1$ by the property $d_{k-1}^* < h$. Therefore, $d_h < \hat{d}_h^* - 1 = \hat{d}_h$, that is, $d_h^+ > d_h$.

THEOREM 20. *Let d be a proper sequence of length n, and let $c \in \mathcal{T}_n$. Then $c \in \mathcal{T}_n(d)$ if and only if $d^- \leqq c \leqq d^+$. Namely, $(\mathcal{T}_n(d), \vee, \wedge)$ is the sublattice of $(\mathcal{T}_n, \vee \wedge)$ induced by the maximum element d^+ and the minimum element d^-.*

Proof. Assume first that $c \in \mathcal{T}_n$ and $d^- \leqq c \leqq d^+$. We may assume that $c \neq d$, for otherwise the theorem is trivial. Let k be the maximum index i such that $d_i^- < c_i$. Such k satisfies the inequalities $m+1 \leqq k \leqq n$ by Lemma 17. Assume for simplicity that $d_k = d_k^+$ (i.e., $\Delta_k < 0$); the case of $d_k = d_k^-$ can be similarly treated. Then one has $d_h^+ > d_h = d_h^-$ for $h = c_k > d_k^-$ by Lemma 19. Moreover it must be $c_h > d_h = d_h^-$, since otherwise $c_k = d_h^-$ for $h = c_k$ (i.e., $c_{h+1} < c_h$) implies $c_k \leqq d_k^-$ by Lemma 19, a contradiction. Consider now the sequence c^1 defined by $c_i^1 = c_i - 1$ for $i = h, k$ and $c_i^1 = c_i$ for $i \neq h, k$; then c^1 is a threshold sequence by Lemma 16, $d^- \leqq c^1 \leqq d^+$ by $c_h > d_h^-$ and $c_k > d_k^-$, and $\|c^1 - d^-\| < \|c - d^-\|$ by $c_h > d_h^-$ and $c_k > d_k^-$. Iterating this procedure, one obtains a finite sequence $c \equiv c^0, c^1, \cdots, c^s \equiv d^-$ of threshold sequences such that $\|c - d\| = \|c^1 - d\| = \cdots = \|d^- - d\|$. Hence $c \in \mathcal{T}_n(d)$.

Assume now that $c \in \mathcal{T}_n$ and, say, $c_i > d_i^+$ for some index i. (The case when $c_i < d_i^-$ can be similarly treated). Let k be the minimum such index. If h is the maximum index of a vertex which is linked to k in the threshold graph with degree sequence c, one must have $c_h > d_h^+$ since k is not linked to h in the threshold graph with degree sequence d^+ (as easily proved by Lemma 9). Therefore the sequence c' defined by $c_i' = c_i - 1$ for $i = h, k$ and $c_i' = c_i$ for $i \neq h, k$ is again threshold by Lemma 16, and $\|c' - d\| < \|c - d\|$. Hence $c \notin \mathcal{T}_n(d)$.

The last half of the theorem then follows immediately from Theorem 13.

Acknowledgments. The support provided by the National Research Council of Canada by a Canada Council Grant for Exchange of Scientists between Japan and Canada and by the Italian National Research Council is gratefully acknowledged.

REFERENCES

[1] C. BENZAKEN AND P. L. HAMMER, *Linear separation of dominating sets in graphs*, Ann. Discrete Math., 3 (1978), pp. 1–10.

[2] C. BERGE, *Graphes et hypergraphes*, Dunod, Paris, 1972.
[3] R. BURKARD AND P. L. HAMMER: *On the Hamiltonicity of split graphs.* University of Waterloo, Department of Combinatorics and Optimization, Research Report CORR 77-40.
[4] V. CHVÁTAL AND P. L. HAMMER, *Aggregation of inequalities in integer programming*, Ann. Discrete Math., 1 (1977), pp. 145-162.
[5] P. ERDÖS AND T. GALLAI, *Graphen mit Punkten vorgeschriebenen Grades*, Mat. Lapok, 11 (1960), pp. 264-274.
[6] S. FÖLDES AND P. L. HAMMER, *On a class of matroid-producing graphs*, Proceedings of the Fifth Hungarian Combinatorial Colloquium, 1976.
[7] ———, *Split graphs having Dilworth number two*, Canad. J. Math., 29 (1977), pp. 666-672.
[8] ———, *Split Graphs*, Eighth Southeastern Conference on Combinatorics, Graph Theory and Computing, 1977.
[9] ———, *The Dilworth number of a graph*, Ann. Discrete Math., 2 (1978), pp. 211-219.
[10] M. C. GOLUMBIC, *Threshold Graphs and Synchronizing Parallel Processes*, Courant Institute of Mathematical Sciences, June, 1976.
[11] P. L. HAMMER AND B. SIMEONE, *The Splittance of a Graph.* University of Waterloo, Department of Combinatorics and Optimization, Research Report CORR 77-39.
[12] P. L. HAMMER, T. IBARAKI AND B. SIMEONE, *Degree sequences of threshold graphs*, Ninth Southeastern Conference on Combinatorics, Graph Theory and Computing, 1978.
[13] F. HARARY, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.
[14] S-Y. R. LI, *Graphic sequences with Unique Realization*, J. Combinatorial Theory (B), 19, (1975), pp. 42-68.
[15] J. ORLIN, *The minimal integral separator of a threshold graph*, Ann. Discrete Math., 1 (1977), pp. 415-419.
[16] C. PAYAN, *Equistable and equidominating graphs*, University of Grenoble, Institute of Applied Mathematics and Informatics, Research Report, 1977.
[17] U. N. PELED, *Matroidal graphs*, Discrete Mathematics, 20 (1977), pp. 263-286.
[18] H. J. RYSER, *Combinatorial Mathematics*, Carus Monographs, American Mathematical Society, Providence RI, 1963.
[19] A. P. WOJDA, *Digraphs of Vicinal Preorder*, Academy of Mining and Metallurgy, Research Report, 1977, Kraków, Poland.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 1, March 1981

0196-5212/81/0201-0007 $01.00/0

A FAST ALGORITHM FOR FINDING STRONG STARTERS*

J. H. DINITZ† AND D. R. STINSON‡

Abstract. A *strong starter* (of order n) in an additive Abelian group G of odd order $n = 2t+1$ is a set $S = \{\{x_1, y_1\}, \{x_2, y_2\}, \cdots, \{x_t, y_t\}\}$ which satisfies the following properties:

(i) $\{x_1, x_2, \cdots, x_t, y_1, y_2, \cdots, y_t\} = G\backslash\{0\}$,

(ii) $\{\pm(y_1 - x_i) | \{x_i, y_i\} \in S\} = G\backslash\{0\}$,

(iii) $x_i + y_i \neq x_j + y_j$ if $i \neq j$, and $x_i + y_i \neq 0$, for any i.

We present a fast algorithm for finding strong starters in Abelian groups.

1. Introduction. Strong starters are used extensively in the construction of Room squares and Howell designs. A *Howell design* $H(n, 2t)$, with $t \leqq n \leqq 2t-1$, is a square array of side n, where cells are either empty or contain an unordered pair of elements chosen from a set X of size $2t$ such that:

(1) each member of X occurs exactly once in each row and column of the array, and

(2) each pair of elements of X occurs in at most one cell of the array.

A *Room square* of side n (n odd) is an $H(n, n+1)$. It follows that, in this case, each pair of elements of X occurs in exactly one cell of the array. Much research has been done concerning Room squares; see, for example [10] and [14]. Strong starters are related to Room squares by the following theorem of Horton [7].

THEOREM 1.1. *If there exists a strong starter of order n, then there exists a Room square of side n.*

Anderson [1], [2] has shown that for the case of Howell designs, the existence of a strong starter of order n which satisfies certain other (technical) properties implies the existence of many $H(n, 2t)$, $(n+1)/2 \leqq t \leqq n$.

For the above reason, strong starters have been investigated by several people. Some infinite classes of strong starters are known. See for example, Mullin and Nemeth [9], Chong and Chan [3], and Gross and Leonard [6]. Indeed, strong starters are known to exist for all orders relatively prime to 3, except for order 5. However, no general method is known for producing strong starters of order $3p$ for p prime. All strong starters of these orders have been found on computer by back-tracking methods (see [4] and [13]). However, for orders exceeding 70, back-tracking becomes impractical due to the excessive computing time required.

Using the algorithm presented in this paper, the authors have recently proven the following theorem [5].

THEOREM 1.2. *If $n < 1000$ is odd, $t \neq n-1$ and $(n, 2t) \neq (5, 6)$, then there exists an $H(n, 2t)$.*

The purpose of this paper is to describe and analyze the algorithm used to find these strong starters.

We wish to point out that we cannot prove that the algorithm will produce a strong starter of any particular order. However, in practice, the algorithm has always succeeded.

In § 2, we describe the algorithm. In § 3, we estimate the time required to be $O(n^2)$ where n is the order n of the strong starter. This estimate agrees with empirical timing results. In § 4, we give a brief geometrical description of strong starters.

* Received by the editors January 30, 1980, and in revised form August 6, 1980.

† Department of Mathematics, University of Vermont, Burlington, Vermont 05405.

‡ Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada.

2. The algorithm. We now present the algorithm used to find a strong starter of order $n = 2t+1$ in the cyclic group $\mathbb{Z}_n$.

Define a *partial strong starter* to be a set $S' = \{\{x_1, y_1\}, \{x_2, y_2\}, \cdots, \{x_r, y_r\}\}$ satisfying the following conditions:

(i) the x_i's and y_i's are distinct nonzero elements of $\mathbb{Z}_n$;

(ii) $y_i - x_i \neq \pm(y_j - x_j)$ if $i \neq j$;

(iii) $x_i + y_i \neq x_j + y_j$ if $i \neq j$, and $x_i + y_i \neq 0$ if $1 \leqq i \leqq r$.

Define def $(S') = t - r$. We say that def (S') is the *deficiency* of S'. The deficiency of S' is the number of "missing pairs". We say that a partial strong starter S' is *maximal* if there exists no $\{u, v\} \subseteq \mathbb{Z}_n$ such that $S' \cup \{\{u, v\}\}$ is a partial strong starter.

In a back-tracking algorithm, when a maximal partial strong starter is reached, the "last" pair $\{x_r, y_r\}$ is deleted from the strong starter. This increases the deficiency of the partial strong starter. The basic feature of the algorithm we will present is that the deficiency is never increased.

Let $D = \{1, 2, \cdots, t\}$. We refer to members of D as *differences*. Then, without loss of generality, we may assume that $y_i - x_i = d_i \in D$, if $1 \leqq i \leqq r$. An element $z \in \mathbb{Z}_n - \{0\}$ is said to be *used* if $z \in \{x_i, y_i\}$ for some $\{x_i, y_i\} \in S'$, otherwise z is *unused*. Similarly, a difference $d \in D$ is said to be *used* or *unused* depending on whether or not $d = d_i$ for some i, $1 \leqq i \leqq r$. Finally, $e \in \mathbb{Z}_n - \{0\}$ is said to be a *used* or *unused sum* depending on whether or not $e = x_i + y_i$ for some i, $1 \leqq i \leqq r$.

We now define a *state* of the algorithm to be a partial strong starter S', together with two distinct unused elements u_1 and u_2, and an unused difference $d \in D$. Given a state of the algorithm, let $T_i = \{u_i - d, u_i + d\}$, $i = 1, 2$, and let $T = T_1 \cup T_2$. The following operations can be performed on a state.

(a) Matching u_i with an unused element:

If there exists $w \in T_i$ such that w is an unused element and $u_i + w$ is an unused sum (for the appropriate $i = 1$ or 2), then let $S'' = S' \cup \{\{u_i, w\}\}$. If def $(S'') \neq 0$, choose a new u_1, u_2, d.

(b) Switching a pair:

If $w \in T_i$ is a used element, and $u_i + w$ is an unused sum, then let $S'' = S' \backslash \{\{x_j, y_j\}\} \cup \{\{w, u_i\}\}$, where $w = x_j$ or y_j for some j, $1 \leqq j \leqq r$. Set

$$d = d_j, \qquad u_1 = u_{3-i},$$

and

$$u_2 = \begin{cases} y_j & \text{if } w = x_j, \\ x_j & \text{if } w = y_j. \end{cases}$$

(c) Back-tracking:

Revert to the previous state of the algorithm if (b) or (c) was the last operation performed.

(d) Switching a difference:

Replace d by some other unused difference d'. Leave u_1, u_2 unchanged.

(e) Switching a pair:

Suppose $u_i - u_{3-i} = d_1 \in D$ is a used difference, and suppose $u_1 + u_2$ is an unused sum. Then set $S'' = S' \backslash \{\{x_{d_1}, y_{d_1}\}\} \cup \{u_1, u_2\}$; set $u_1 = x_{d_1}$, $u_2 = y_{d_1}$, and leave d unchanged.

We may now use operations (a)–(e) to describe our algorithm.

(1) Initialization: Set def $= t$, $S = \varnothing$, choose any distinct $u_1, u_2 \in \mathbb{Z}_n - \{0\}$, $d \in D$.

(2) If operation (a) can be performed, do so and go to (8).
(3) If operation (b) can be performed, do so and go to (2).
(4) If operation (c) can be performed, do so and go to (3).
(5) If operation (d) can be performed, do so and go to (2).
(6) If operation (e) can be performed, do so and go to (2).
(7) Stop (algorithm fails).
(8) Set def = def − 1, choose any distinct unused u_1, u_2 and d. If def $\neq 0$ go to (2).
(9) Stop (algorithm succeeds).

A few comments regarding the algorithm are in order. First, no operation increases the deficiency, and operation (a) decreases the deficiency by 1. Also, operations (d) and (e) are rarely executed since it is unlikely that (a), (b) and (c) all fail (more details in § 3). Note that if def = 1, then operation (d) cannot occur, since (d) requires an unused difference other than d. Finally, note that there may be more than one way to perform an operation (b) on a given state. As a heuristic in the implementation of the algorithm, the following is done. If a state is reached, and more than one way to perform operation (b) is possible, then one way is picked at random. If the state is again reached, this time by back-tracking (operation (c)), then the first way to perform operation (b) is excluded and one of the remaining ways is chosen at random. As an example of this, see lines 9–12 in Table 1 below.

We construct a strong starter of order 11 using this algorithm. Table 1 below traces the execution of the algorithm. Note that no operations (d) or (e) were required.

TABLE 1

Partial strong starter					State			Operation to be performed
diff 1	2	3	4	5	u_1	u_2	d	
					9	5	3	a
		1, 9			2	3	5	a
		1, 9		7, 2	3	4	4	a
		1, 9	3, 10	7, 2	4	5	2	b
	4, 2	1, 9	3, 10		5	7	5	b
	4, 2		3, 10	1, 7	5	9	3	a
	4, 2	9, 6	3, 10	1, 7	5	8	1	b
5, 4		9, 6	3, 10	1, 7	8	2	2	b
5, 4	10, 8	9, 6		1, 7	2	3	4	b
5, 4	10, 8	9, 6	7, 3		2	1	5	c
5, 4	10, 8	9, 6		1, 7	2	3	4	c
5, 4		9, 6	3, 10	1, 7	8	2	2	b
5, 4	8, 6		3, 10	1, 7	2	9	3	b
5, 4	8, 6	1, 9	3, 10		2	7	5	c
5, 4	8, 6		3, 10	1, 7	2	9	3	b
	8, 6	5, 2	3, 10	1, 7	9	4	1	b
9, 8		5, 2	3, 10	1, 7	4	6	2	a
9, 8	6, 4	5, 2	3, 10	1, 7				

3. Analysis of the algorithm. In this section, we estimate the efficiency of this algorithm by some probabilistic considerations and present some empirical data.

In order to calculate this estimate, one major assumption is made. We assume that the probability that an operation succeeds on a given state is independent of the

previous state performed. Theoretically, this assumption is probably not even true. However, analysis of the efficiency of this algorithm using the assumption of independence strongly agrees with the empirical data (see Tables 2 and 3). It thus appears (as

TABLE 2

100 strong starters of each order n						
n	Average of $N_a+N_b+N_c+N_d+N_e$	Average $\frac{}{n \log n}$	95% Confidence interval	Average of N_d+N_e	Number that failed	Time[1]
51	393	1.95	53	.39	17	6.59 sec
101	757	1.62	96	.89	13	9.29
201	1611	1.51	173	2.05	11	18.21
301	2491	1.45	272	3.31	12	29.32
401	3486	1.45	317	4.64	10	41.35
501	4029	1.29	374	5.29	10	51.02
601	4932	1.28	423	7.16	12	65.25

TABLE 3

2 strong starters of each order n				
n	Average of $N_a+N_b+N_c+N_d+N_e$	Average $\frac{}{n \log n}$	Time	$\frac{\text{Time}^1}{n^2}\times 10^6$
3001	31190	1.29	14.4 sec	1.60
5001	63645	1.50	32.0	1.28
8001	91852	1.28	77.5	1.21
10001	117020	1.27	117.1	1.17

intuition would indicate) that the states are nearly independent, particularly for n large. Because of the independence assumption, the analysis which follows is merely an estimate of the actual efficiency of the algorithm and is not a proof of the existence of strong starters.

First we estimate the probability that operation (a) succeeds for a given state with deficiency k. The number of unused elements, other than u_1 or u_2, is $2k-2$. If operation (b) was just performed, then one element of T will be used. The other three elements of T each have probability $(2k-2)/n$ of being unused and distinct from u_1 and u_2. The probability that a given element of T is unused is less than the probability p that there is some unused element in T. Thus, for some element $e \in T$, distinct from u_1 and u_2, the probability that e is unused is $(2k-2)/(n-2)$. So a lower bound on p is $p \geqq (2k-2)/(n-2)$.

There is also the possibility that $u_1-u_2=\pm d$. This happens with probability $2/(n-1)$. Finally, the probability that a given sum is nonzero and unused is $((n-1)/2+$

[1] The algorithm was implemented in Fortran on The Ohio State University Amdahl 470 system.

$k)/n$. Thus, the probability of (a) succeeding when the deficiency is k is at least

$$\begin{aligned} p_k(\mathrm{a}) &> \left[\binom{2k-2}{n-2} + \frac{2}{n-1}\right]\frac{n+2k-1}{2n} \\ &> \binom{2k-2+2}{n-1}\binom{n+2k-1}{2n} \\ &= \frac{k(n+2k-1)}{(n-1)n}. \end{aligned}$$

Thus, the number N_{a} of times operation (a) is attempted in the course of the algorithm is approximately

$$\begin{aligned} \sum_{k=1}^{(n-1)/2} \frac{1}{p_k(\mathrm{a})} &< (n-1)n \sum_{k=1}^{(n-1)/2} \frac{1}{k(n+2k-1)} \\ &< n \sum_{k=1}^{(n-1)/2} \frac{1}{k} \\ &< n\left(\log\left(\frac{n-1}{2}\right)+1\right). \end{aligned}$$

Thus, it appears that $N_{\mathrm{a}} = O(n \log n)$.

Now, if we suppose that (c) does not fail in the course of the algorithm, we can have that $N_{\mathrm{a}} = N_{\mathrm{b}} - N_{\mathrm{c}^*} + (n-1)/2$. N_{b} and N_{c^*} denote the number of times operations (b) and (c*) are attempted, where an operation (c*) is a maximal sequence of consecutive operations (c).

We now compute the probability $p_k(\mathrm{b})$ of (b) succeeding if (a) or (b) was just performed. If (a) was just performed, then there are four possibilities for pairs to be switched, if (b), then three. The probability of at least one sum being unused is at least

$$1-\left(\frac{(n-1)/2-k}{n}\right)^3 = 1-\left(\frac{n-2k-1}{n}\right)^3 > \frac{7}{8}.$$

If w were an unused element, then (a) would be performed. Thus, w is used (perhaps zero). If w is nonzero then (b) can be performed. In order to simplify the arithmetic, we assume w is nonzero. This does not greatly affect our estimate. Thus we estimate $p_k(\mathrm{b}) > \frac{7}{8}$. Since (c) occurs only after (b) fails, we have $N_{\mathrm{c}^*} < \frac{1}{8}N_{\mathrm{b}}$.

Finally, the number of operations (c) in one operation (c*) must be estimated. Denote by $p_k(\mathrm{b}=1)$ the probability that there is exactly one permissible choice in a (b) operation (where a (b) or (a) was just performed). Then

$$\begin{aligned} 1-p_k(\mathrm{b}=1) &= 1-3\left(\frac{n-2k+1}{2n}\right)^2\left(\frac{n+2k-1}{2n}\right) \\ &\geqq 1-3\left(\frac{n-1}{2n}\right)^2\left(\frac{n+1}{2n}\right) \\ &\geqq \frac{5}{8}. \end{aligned}$$

Thus we estimate that, on the average, less than $\frac{8}{5}$ (c) operations make up each (c*) operation.

By the above, we have $N_b = N_a + N_{c^*} - (n-1)/2$. Thus $N_b < N_a + \frac{1}{8}N_b - (n-1)/2$, so $N_b < \frac{8}{7}(N_a - (n-1)/2)$. Therefore, $N_b = O(n \log n)$. Also, $N_c < \frac{8}{5}N_{c^*} < \frac{1}{5}N_b$, so $N_c = O(n \log n)$.

Thus, we estimate that the number of operations, $N_a + N_b + N_c$, executed in the algorithm is $O(n \log n)$. Also, the time required for an operation is at most $O(n)$. Choosing a new u_1 and u_2 is the only time it is necessary to search through an array. With more sophisticated list processing techniques, this time could be reduced, perhaps to $O(\log n)$. Each of the operations (a)–(e) require $O(1)$ time. Thus, we estimate that the time required for the algorithm is $O(n^2 \log n)$.

This estimate can be improved slightly. The $O(n)$ operation is executed only $(n-1)/2$ times in the course of the algorithm. Thus, an estimate of $O(n^2)$ is obtained.

To test this estimate, the program was run until 100 strong starters were produced for each of 7 different orders, (See Table 2). Also, a 95% confidence interval about the mean μ of $N_a + N_b + N_c + N_d + N_e$ was computed. To test the algorithm on large orders, we produced two strong starters each of orders 3001, 5001, 8001, and 10001 (Table 3).

Although no theoretical upper bound for the probability of failure has been computed, in practice this number appears to be about $\frac{1}{10}$. The algorithm usually fails when deficiency equals 1 and no operation can be performed. This happens only in the first state after the deficiency has become 1, since otherwise the program will be able to back-track when no (b) can be performed. There is also the chance that the states might form a loop and thus not produce a starter. In order to prevent an infinite loop, a timer was written into the program. If the search for a starter took too long, the search would be aborted and the program started over again with def $= (n-1)/2$. However, this occurred only once in over 700 trials.

4. A geometric interpretation. Strong starters in $\mathbb{Z}_n$ have an interesting geometrical interpretation. Label n equally spaced points on a circle by the elements of $\mathbb{Z}_n$ (cyclically). If $\{x, y\} \in S$, then join points x and y on the circle by a straight line. The $(n-1)/2$ lines thus formed will have the following properties:

(1) no two lines have the same length;

(2) no two lines are parallel;

(3) no two lines have a common endpoint.

Conversely, any such geometric configuration generates a strong starter in $\mathbb{Z}_n$.

A strong starter of order 129 is geometrically represented in Fig. 1 below.

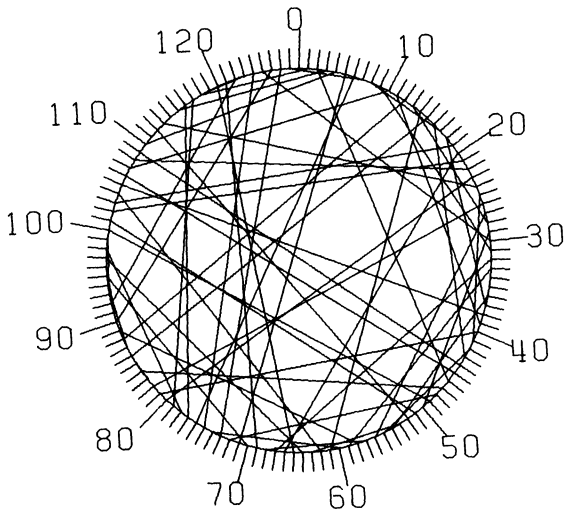


FIG. 1. *A strong starter of order* 129.

5. Conclusion. Thus, we have described an algorithm for finding strong starters. Using probabilistic arguments, we estimate that the algorithm should succeed in polynomial time (actually $O(n^2)$). In practice, this seems to be accurate.

Our algorithm is similar in some aspects to the algorithm of Posa [11] for finding Hamiltonian circuits in graphs. That is, at no time in the algorithm does one head "away" from the desired end results. In finding strong starters, the deficiency is never increased; in finding a Hamiltonian circuit, the length of a path is never decreased. Also, both algorithms involve a certain amount of randomness in making some choices. Finally, there is the possibility that the algorithm may fail. However, in practical applications both algorithms have a high rate of success.

Other probabilistic algorithms are described in [8] and [12].

It also appears that probabilistic algorithms based on this simple switching idea may be practical in other combinatorial applications such as constructing Steiner triple systems and finding transversals in Latin squares.

REFERENCES

[1] B. A. ANDERSON, *Howell designs from Room squares*, Proc. 2nd Caribbean Conf. on Comb. and Comp., Barbados, 1977, pp. 55–62.

[2] ———, *Starters, digraphs, and Howell designs*, Utilitas Math., 14 (1978), pp. 219–248.

[3] B. C. CHONG AND K. M. CHAN, *On the existence of normalized Room squares*, Nanta Math., 7 (1974), pp. 8–17.

[4] R. J. COLLENS AND R. C. MULLIN, *Some properties of Room squares – a computer search*, Proc. 1st Louisiana Conf. on Combinatorics, Graph Theory and Computing, Baton Rouge, 1970, pp. 87–111.

[5] J. H. DINITZ AND D. R. STINSON, *A note on Howell designs of odd side*, Utilitas Math., to appear.

[6] K. B. GROSS AND P. A. LEONARD, *The existence of strong starters in cyclic groups*, Utilitas Math., 7 (1975), pp. 187–195.

[7] J. D. HORTON, *Room designs and one-factorizations*, Aequationes Math., to appear.

[8] R. M. KARP, *The probabilistic analysis of some combinatorial search algorithms*, in Algorithms and Complexity, Academic Press, New York, 1976.

[9] R. C. MULLIN AND E. NEMETH, *An existence theorem for Room squares*, Canad. Math. Bull., 12 (1969), pp. 493–497.

[10] R. C. MULLIN AND W. D. WALLIS, *The existence of Room squares*, Aequationes Math., 13 (1975), pp. 1–7.

[11] L. POSA, *Hamilton circuits in random graphs*, Discrete Math., 14 (1976), pp. 359–364.

[12] M. O. RABIN, *Probabilistic Algorithms*, in Algorithms and Complexity, Academic Press, New York, 1976.

[13] R. G. STANTON AND R. C. MULLIN, *Construction of Room squares*, Ann. Math. Statist., 39 (1968), pp. 1540–1548.

[14] W. D. WALLIS, A. P. STREET AND J. S. WALLIS, *Combinatorics: Room Squares, Sum-free Sets, Hadamard Matrices*, Lecture Notes in Mathematics 292, Springer-Verlag, Berlin, 1972.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 1, March 1981

0196-5212/81/0201-0008 $01.00/0

DIAGONAL SCALING TO AN ORTHOGONAL MATRIX*

A. BERMAN†, B. N. PARLETT‡ AND R. J. PLEMMONS§

Abstract. An algorithm is given which determines whether a matrix A is diagonally equivalent to an orthogonal matrix and, if so, computes the corresponding scaling factors. The algorithm makes use of the Hadamard quotient $A^{-1} \oplus A^t$. Such problems arise, for example, in the study of energy conserving norms for the solution of hyperbolic systems of partial differential equations.

1. Introduction. The problem of finding an energy conserving norm for the solution of the hyperbolic system of partial differential equations $\partial u/\partial t = W\, \partial u/\partial x$, with $t>0$ and $0<x<1$ and where W is diagonalizable, subject to boundary conditions, has been reduced by Gunzburger and Plemmons [1979] to the problem of characterizing those matrices S_1 and S_2, appearing in the boundary conditions, which enjoy the following properties:

(i) S_1 and S_2 are invertible;

(ii) there exist positive diagonal matrices D and E such that DS_1E and $DS_2^{-1}E$ are orthogonal matrices.

This characterization problem is completely solved, in the paper cited above, only for matrices of order $n \leqq 2$. To deal with the cases when $n>2$, we study the problem of determining when it is possible to row and column scale a real square matrix A to produce an orthogonal matrix. If D and E are positive diagonal matrices and

$$DAE = Q, \qquad Q \text{ orthogonal},$$

then we say that A and Q are *diagonally equivalent* and A is d.e.o. (*diagonally equivalent to an orthogonal matrix*).

This paper gives a necessary and sufficient condition for A to be d.e.o. Of more importance, it offers an algorithm which is constructive in the sense that either it yields D and E such that DAE is orthogonal, or it fails and no such pair D, E exists. The presentation confines itself to real matrices but all the results extend in the standard way to the scaling of a complex matrix into a unitary one.

Clearly, A is d.e.o if and only if A^{-1} and A^T are diagonally equivalent. Thus, the more general diagonal equivalence theorems given by Sinkhorn and Knopp [1969] and by Engel and Schneider [1973], [1975], could essentially be applied to our case. However, some modifications of their algorithms would probably be necessary to make them competitive on the specific problem addressed here. Further work on the general diagonal equivalence problem has been reported by Engel and Schneider [1980], where results in Saunders and Schneider [1978] are improved and extended.

2. The Hadamard quotient. The *Hadamard product* (or Schur product) $F \odot G$ of two matrices, F and G, of the same size, is defined by $(F \odot G)_{ij} = f_{ij}g_{ij}$ and occurs in various parts of matrix theory. Let $H(A)$ denote any $\binom{n}{2} \times n$ matrix whose rows are the Hadamard products of pairs of distinct rows of A. It is not difficult to see that A is d.e.o.

* Received by the editors March 17, 1980 and in revised form July 7, 1980.

† Department of Mathematics, Technion-Israel Institute of Technology, Haifa 32000, Israel, and Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York 12181.

‡ Departments of Mathematics and the Computer Science Division of the EECS Department, University of California, Berkeley, California 94720. The research of this author was supported in part by the U.S. Office of Naval Research under contract N00014-76-C-0013.

§ Departments of Computer Science and Mathematics, University of Tennessee, Knoxville, Tennessee 37916. Research supported in part by the U.S. Army Research Office under grant DAAG29-80-K-0025.

if and only if the system $H(A)x = 0$ has a positive solution x. An important consequence is that A is not d.e.o. if it has two distinct rows (or columns, since A^T is d.e.o. if and only if A is d.e.o.) whose Hadamard product is nonzero and nonnegative or nonpositive. However, computationally, the Hadamard product is not as useful in our scaling problems as is the less well-known element-by-element quotient (*Hadamard quotient*) of two matrices. Formally, we define $H = F \oslash G$ by $h_{ij} = f_{ij}/g_{ij}$ and we allow elements of H to be either infinite (1/0) or undefined (0/0). We use the *symbol u for the column vector whose elements are all* 1. The square matrix of 1's has rank one and may be written as uu^T.

The Hadamard quotient permits a strange but fruitful formulation of the definition of an orthogonal matrix, namely that A is orthogonal if and only if the well-defined elements of $Q^{-1} \oslash Q^T$ are all 1. In order to exploit this approach, we introduce two bits of terminology. For any invertible matrix B, we write

$$\Phi(B) \equiv B^{-1} \oslash B^T. \tag{2.1}$$

Next, in order to cope smoothly with undefined elements in $\Phi(B)$ we define a partial equality (not a true equivalence relation) on matrices by

$$A \triangleq B \quad \text{if } a_{ij} = b_{ij}, \quad \text{whenever } a_{ij} \text{ and } b_{ij} \text{ are well defined or infinite.}$$

With these notions in hand, our formulation of orthogonality becomes

$$Q \text{ is orthogonal} \quad \text{iff} \quad \Phi(Q) \triangleq uu^T, \quad \text{that is, a matrix of all 1's.} \tag{2.2}$$

Our algorithm was suggested by the simple effect of scaling on the very nonlinear function Φ. If $D = \operatorname{diag}(d_1, \cdots, d_n)$ and $E = \operatorname{diag}(e_1, \cdots, e_n)$ are invertible scaling matrices, then it is easy to verify that

$$\Phi(DAE) = E^{-2}\Phi(A)D^{-2}. \tag{2.3}$$

This relation allows us to characterize matrices which are d.e.o. (diagonally equivalent to an orthogonal matrix).

THEOREM 1. *An invertible matrix A is* d.e.o. *if and only if* $\Phi(A) \triangleq$ *a positive, rank one matrix.*

Proof.

$$\begin{aligned} \text{DAE is orthogonal} &\Leftrightarrow \Phi(DAE) \triangleq uu^T, \quad \text{by (2.2)}, \\ &\Leftrightarrow E^{-2}\Phi(A)D^{-2} \triangleq uu^T, \quad \text{by (2.3)}, \\ &\Leftrightarrow \Phi(A) \triangleq (E^2u)(D^2u)^T. \end{aligned}$$

Thus,

$$(\Phi(A))_{ij} = \begin{cases} 0/0 & \text{if } a_{ji} = (A^{-1})_{ij} = 0, \\ e_i^2 d_j^2 \ (>0) & \text{otherwise.} \end{cases}$$

□

The proof reveals how D and E must be chosen when $\Phi(A)$ is a positive rank one matrix. Of course, D and E themselves are not unique because the product xy^T does not determine x and y uniquely.

Remark. Two d.e.o. matrices A and B, for example S_1 and S_2^{-1} in the energy conserving norm example that motivated our discussion, can be scaled simultaneously if and only if $\Phi(A) \triangleq \Phi(B)$.

3. A simple algorithm. For the sake of clarity, we will specify here a scaling algorithm for a special class of matrices, namely $n \times n$ matrices which have at least one column, say the kth, which has no zero elements.

Specification. The algorithm sets DEO to the value *true* and returns the diagonal elements of D and E if A is d.e.o. Otherwise, DEO is given the value *false* and D and E are left undefined.

Algorithm 1.

1. DEO $\leftarrow$ false.
2. Compute A^{-1}. If inversion fails, go to 8.
3. Compute $C = \Phi(A)$. If C has any nonpositive or infinite elements, go to 8. Record undefined elements (0/0) as 0's.
4. $d_j \leftarrow c_{kj}$, $j = 1, \cdots, n$. (The kth column of A is assued to have no zero elements.)
5. $e_i \leftarrow c_{iq}/d_q$, for any positive element c_{iq}, $i = 1, \cdots, n$.
6. Check the rank one property: If ($c_{ij} \neq 0$ and $c_{ij} \neq e_i d_j$), $i, j = 1, \cdots, n$, go to 8.
7. $d_j \leftarrow \sqrt{d_j}$, $e_j \leftarrow \sqrt{e_j}$, $j = 1, \cdots, n$. DEO $\leftarrow$ true.
8. Exit.

Example.

$$A = \begin{bmatrix} 1 & 1 & \sqrt{3} \\ -1 & 1 & \sqrt{3} \\ 0 & -\sqrt{6} & \sqrt{2} \end{bmatrix}, \qquad A^{-1} = \frac{1}{8}\begin{bmatrix} 4 & -4 & 0 \\ 1 & 1 & -\sqrt{6} \\ \sqrt{3} & \sqrt{3} & \sqrt{2} \end{bmatrix},$$

$$C = \Phi(A) = \frac{1}{8}\begin{bmatrix} 4 & 4 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \doteq \begin{bmatrix} 4 \\ 1 \\ 1 \end{bmatrix} \left(\tfrac{1}{8}, \tfrac{1}{8}, \tfrac{1}{8}\right).$$

Here, we can take

$$D = \frac{1}{2\sqrt{2}} I, \qquad E = \operatorname{diag}(2, 1, 1).$$

Of course,

$$D = I, \qquad E = \operatorname{diag}\left(\frac{1}{\sqrt{2}}, \frac{1}{2\sqrt{2}}, \frac{1}{2\sqrt{2}}\right)$$

also suffice. Note that the undefined $(1, 3)$ element of $\Phi(A)$ does not impair the formation of D and E.

4. Canonical permutations. In the general case, it is not obvious how to pick the right values for d_j in step 4 for $D = \operatorname{diag}(d_1, \cdots, d_n)$, when C has many zero elements. In order to be able to deal with this case, we consider the effect of permutations. Let P_1 and P_2 denote permutation matrices.

LEMMA 1. *A is* d.e.o. *if and only if $P_1 A P_2$ is* d.e.o.

Proof. Since each P_i is orthogonal, we see that

$$\begin{aligned} \text{DAE is orthogonal} &\Leftrightarrow P_1(DAE)P_2 \text{ is orthogonal,} \\ &\Leftrightarrow (P_1 D P_1^T)(P_1 A P_2)(P_2^T E P_2) \text{ is orthogonal,} \\ &\Leftrightarrow \hat{D}(P_1 B P_2)\hat{E} \text{ is orthogonal,} \end{aligned}$$

where $\hat{D} = P_1 D P_1^T$ and $\hat{E} = P_2^T E P_2$ are also invertible diagonal matrices. □

It follows that there is no loss in applying the full power of permutations to simplify A. In particular, when A has many zero elements, its inversion can be done more quickly if permutations are used adroitly. An important tool in the following discussion is a canonical form under two-sided permutations. Recall that a square matrix B is *indecomposable* if there exists no permutation matrix P such that

$$PBP^T = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}, \qquad A_{ii} \text{ square}, \tag{4.1}$$

and *fully indecomposable* if no permutation matrices P_1 and P_2 exist such that P_1BP_2 has the form (4.1).

Several references to the following theorem are given, for example, in Duff [1977] and Howell [1976].

THEOREM 2. *Given a square nonsingular matrix A, there exist permutation matrices P_1 and P_2 such that*

$$P_1AP_2 = \begin{bmatrix} A_{11} & & & \\ A_{21} & A_{22} & & 0 \\ \vdots & \vdots & \ddots & \\ A_{r1} & A_{r2} & \cdots & A_{rr} \end{bmatrix}, \tag{4.2}$$

where each A_{ii} is square nonsingular and fully indecomposable and its diagonal elements are nonzero.

Algorithms for permuting A into the form (4.2) can be found, among other places, in Howell [1976] and in Duff and Reid [1978]. An efficient computer program for permuting A into the form (4.2) is included in the HARWELL Sparse Matrix Subroutine Package MA28 (see Duff and Reid [1979]).

If the matrix in (4.2) is orthogonal, then it is easy to see that $A_{ij} = 0$, $i > j$, and each A_{ii} is orthogonal. Since diagonal scaling does not affect the zero pattern of a matrix, it follows that if A is d.e.o. then the matrix in (4.2) is block diagonal.

5. The general case. The general case is thus reduced to considering (fully) indecomposable A's with nonzero diagonals. Such A's have a special property which is useful in the present context. We say that A has a covering sequence of overlapping columns (c.s.o.c.), $A^{j_1}, \cdots, A^{j_m}$, $j_1 < \cdots < j_m$, if:

a) For each i, $1 \leqq i \leqq n$, there exists q such that $a_{ij_q} \neq 0$. This "covers" the set $\{1, \cdots, n\}$.

b) Each column has a nonzero position in common with some earlier column; i.e., given $q > 1$ there is a $p < q$ and an index i such that $a_{ij_p} \cdot a_{ij_q} \neq 0$. This is the overlap.

In the example below, where X denotes a nonzero element, columns 2, 3 and 5 form such a sequence but 1, 2, 3, 4, 5, 6 do not.

$$A = \begin{bmatrix} X & 0 & 0 & X & X & X \\ 0 & X & 0 & 0 & 0 & X \\ 0 & X & X & X & 0 & 0 \\ X & 0 & X & X & 0 & 0 \\ 0 & 0 & X & X & X & 0 \\ 0 & 0 & 0 & 0 & X & X \end{bmatrix}.$$

To show that the diagonal blocks in (4.2) enjoy the c.s.o.c. property, we prove the following:

LEMMA 2. *If all the diagonal elements of an indecomposable matrix A are nonzero, then A has a covering sequence of overlapping columns.*

Proof. Let α denote a maximal set of indices "covered" by a sequence of overlapping columns. The adjective maximal qualifies $|\alpha|$, the number of distinct indices in α.

If $|\alpha| < n$, consider a_{ij}, $i \in \alpha$, $j \notin \alpha$. If $a_{ij} \neq 0$, the column j "overlaps" column i because $a_{ii} \neq 0$. Since the diagonal elements are nonzero, α contains the indices of all columns in the covering sequence. Thus j could be put into α, thereby increasing $|\alpha|$ and contradicting α's maximality. Consequently, $a_{ij} = 0$ for all $i \in \alpha$, $j \notin \alpha$. This contradicts the assumption that A is indecomposable and it follows that $|\alpha| = n$. Thus, any maximal overlapping sequence of columns covers $\{1, \cdots, n\}$. □

An equivalent formulation of Lemma 2 is that a fully indecomposable matrix has a c.s.o.c. This follows from the fact that the c.s.o.c property is invariant under permutation of rows and from the well-known Brualdi, Parter and Schneider [1966] result that A is fully indecomposable if and only if PA is indecomposable and all its diagonal elements are nonzero for some permutation matrix P. In the case when all the diagonal elements of A are nonzero, full indecomposability is equivalent to indecomposability. Notice that we did not use the invertibility of the diagonal blocks under discussion.

In Algorithm 1, a covering sequence consisted of a single column. The existence of a covering sequence of overlapping columns of A permits the construction of an appropriate *scaling vector d* from $\Phi(A)$, despite the presence of undefined elements in every row.

Construction of d. Without loss of generality, let the indices of a covering sequence of columns of A be $1, \cdots, m$, $(m < n$ if $n > 1)$. As in Algorithm 1, if $C = \Phi(A)$ contains any nonpositive or infinite elements, then the construction fails and A is not d.e.o. Otherwise, for $i = 1, \cdots, m$ repeat:

a) If $i = 1$ set $\rho = 1$, otherwise pick $q < i$ and j, such that $c_{qj} > 0$, $c_{ij} > 0$, and set $\rho = c_{qj}/c_{ij}$.

b) Set $d_k = c_{ik}\rho$ for each k such that $c_{ik} > 0$. If this operation changes any previously assigned (positive) value of a d_k, then A is not d.e.o.

Our construction insures that c_{ij} and c_{qj} yield the same value for d_j. The overlapping property insures that a j exists in step (a) for some $q < i$. The covering property insures that all elements of d are specified when the construction terminates.

6. The algorithm.

Specification. Given an $n \times n$ matrix A, Algorithm 2 either sets DEO to *true* and computes the diagonal elements of D and E so that DAE is orthogonal or, if no such pair D, E exists, it sets DEO to *false* and leaves D and E undefined.

Algorithm 2.

1. DEO ← false.
2. If A has a covering sequence of overlapping columns, then $r \leftarrow 1$, set $A_{11} = A$, go to 4.
3. Put A into a canonical form (4.2) as shown in Theorem 2. If any off-diagonal block is nonnull go to 12.
4. Repeat steps 5–10 for $\mu = 1, \cdots, r$.
5. Compute $A_{\mu\mu}^{-1}$. If inversion fails, go to 12.
6. Compute $C^{\mu} = \Phi(A_{\mu\mu})$. If C^{μ} has any nonpositive or infinite elements go to 12. Record undefined elements of C^{μ} as 0.
7. Find a covering sequence of overlapping columns of $A_{\mu\mu}$ and use the construction in § 5 to define d^{μ}. If the construction fails, go to 12.

8. Define e^{μ} via $e_i^{\mu} \leftarrow c_{iq}^{\mu}/d_q^{\mu}$ for any positive c_{iq}^{μ}.
9. Check that $C^{\mu} \stackrel{?}{=} e^{\mu} d^{\mu T}$; i.e., for i, j ranging over the indices of C^{μ} execute: if $c_{ij}^{\mu} \neq 0$ and $c_{ij}^{\mu} \neq e_i^{\mu} d_j^{\mu}$ then go to 12.
10. $d_j^{\mu} \leftarrow \sqrt{d_j^{\mu}}$, $e_j^{\mu} \leftarrow \sqrt{e_j^{\mu}}$ for all j belonging to C^{μ}.
11. Rearrange e and d^T by undoing the permutations used to achieve the canonical form (see Lemma 1). Then set $D = \operatorname{diag}(d_1, \cdots, d_n)$, $E = \operatorname{diag}(e_1, \cdots, e_n)$ and DEO $\leftarrow$ true.
12. Exit.

We conclude our discussion with an example and some remarks.

Example. Let

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & \sqrt{3} & 0 & 0 \\ 0 & -1 & 1 & 0 & \sqrt{3} & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & -\sqrt{6} & 0 & \sqrt{2} & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & -1 \end{bmatrix}.$$

Then since each column of A contains zeros, Algorithm 1 cannot be applied. Moreover, A has no covering sequence of overlapping columns. From applying step 3 of Algorithm 2, it follows that a canonical form (4.2) for A is

$$PAP^T = \begin{bmatrix} 1 & 0 & 1 & -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & \sqrt{3} \\ 0 & 0 & 0 & 0 & -1 & 1 & \sqrt{3} \\ 0 & 0 & 0 & 0 & 0 & -\sqrt{6} & \sqrt{2} \end{bmatrix},$$

where P is the permutation matrix representing the index permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 4 & 5 & 6 & 2 & 7 & 1 & 3 \end{pmatrix}.$$

Software for permuting general matrices into the form (4.2) is readily available (see Duff and Reid [1979]). Observing that no off-diagonal block is nonnull, let

$$A_{11} = \begin{bmatrix} 1 & 0 & 1 & -1 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}.$$

Then

$$A_{11}^{-1} = \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & -1 & 1 & 1 \\ 1 & 0 & -1 & 1 \\ -1 & 1 & 0 & 1 \end{bmatrix},$$

and

$$C^1 = \Phi(A_{11}) = \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}.$$

Here, as before, we have recorded undefined terms $0/0$ in C^1 as zeros. Observe that 1, 2 are the indices of a covering sequence of overlapping columns of A_{11}. Thus, we proceed to the construction in § 5 to define d^1. For $i = 1$ we set $\rho = 1$ and have $d_1 = c_{11} = \frac{1}{3}$, $d_2 = c_{12} = \frac{1}{3}$ and $d_3 = c_{13} = \frac{1}{3}$. For $i = 2$, we pick $q = 1$ and $j = 2$ so $\rho = c_{12}/c_{22} = 1$. Then, again, $d_2 = c_{22} = \frac{1}{3}$, $d_3 = c_{23} = \frac{1}{3}$, and $d_4 = c_{24} = \frac{1}{3}$. Thus, $d^1 = \text{diag}\,(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Returning to step 8 of Algorithm 2, we see that $e^1 = \text{diag}\,(1, 1, 1, 1)$. Then

$$C^1 \triangleq e^1 d^{1T},$$

and A_{11} is d.e.o. From step 10,

$$D_1 = \frac{1}{\sqrt{3}} I, \qquad E_1 = I.$$

Next, observe that

$$A_{22} = \begin{bmatrix} 1 & 1 & 3 \\ -1 & 1 & 3 \\ 0 & -6 & 2 \end{bmatrix}$$

is the same matrix A as in the example of § 3. Thus, Algorithm 1 applies here and, as before, A_{22} is d.e.o. with

$$D_2 = \frac{1}{2\sqrt{2}} I, \qquad E_2 = \text{diag}\,(2, 1, 1).$$

Next, we undo the permutations yielding

$$D = P^T \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} P = \text{diag}\left(\frac{1}{\sqrt{3}}, \frac{1}{2\sqrt{2}}, \frac{1}{2\sqrt{2}}, \frac{1}{\sqrt{3}}, \frac{1}{2\sqrt{2}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right),$$

$$E = P^T \begin{bmatrix} E_1 & 0 \\ 0 & E_2 \end{bmatrix} P = \text{diag}\,(1, 2, 1, 1, 1, 1, 1, 1).$$

Finally, the orthogonal matrix Q which is diagonally equivalent to A by D and E is given by

$$Q = DAE = \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 & 0 & \frac{1}{\sqrt{3}} & 0 & 0 & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{2\sqrt{2}} & 0 & \frac{\sqrt{3}}{2\sqrt{2}} & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{2\sqrt{2}} & 0 & \frac{3}{2\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{3}} & 0 & 0 & -\frac{1}{\sqrt{3}} & 0 & \frac{1}{\sqrt{3}} & 0 \\ 0 & 0 & -\frac{6}{2\sqrt{2}} & 0 & \frac{1}{2} & 0 & 0 \\ -\frac{1}{\sqrt{3}} & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ 0 & 0 & 0 & \frac{1}{\sqrt{3}} & 0 & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \end{bmatrix}.$$

Remarks.

1. The algorithm needs an extra $n \times n$ array, besides A, to hold A^{-1} and then C. It also requires n-vectors for D, E and the permutations P_1 and P_2. It is not strictly necessary to compute $C = \Phi(A)$ but it is convenient.

2. In practice, the relation $c_{ij} \neq e_i d_j$ would be replaced by $|c_{ij} - e_i d_j| > \varepsilon |c_{ij}|$ for some suitable small ε depending on the precision of the arithmetic unit. Likewise, elements α of A^{-1} should be taken as zero if $|\alpha| < \varepsilon \|A^{-1}\|$.

3. If A has a covering sequence of overlapping columns, e.g., if A has a column with no zeros, then A must be fully indecomposable in order to be diagonally equivalent to an orthogonal matrix. We have not provided an algorithm for finding such a sequence when it exists. However, Professor Hans Schneider has suggested that he has in mind such a construction, which could also be used as an alternate proof to Lemma 2. His method, however, may produce redundant columns in the sequence. Finally, M. T. Heath has suggested an algorithm based upon the bipartite graph of a matrix A which will usually determine a minimal covering sequence of overlapping columns in an efficient manner, whenever A is indecomposable with no zero diagonal entries. Further work is needed on this topic.

Acknowledgment. The authors wish to thank Professor Hans Schneider for his comments on the general diagonal equivalence problem.

REFERENCES

R. A. BRUALDI, S. V. PARTER AND H. SCHNEIDER [1966], *The diagonal equivalence of a nonnegative matrix to a stochastic matrix*, J. Math. Anal. Appl., 16, pp. 31–50.

I. S. DUFF [1977], *A survey of sparse matrix research*, Proc. IEEE, 65, pp. 500–525.

I. S. DUFF AND J. K. REID [1978], *An implementation of Tarjan's algorithm for the block triangularization of a matrix*, ACM Trans. Math. Software, 4, pp. 137–147.

——— [1979], *Some design features of a sparse matrix code*, ACM Trans. Math. Software, 5, pp. 18–35.

G. M. ENGEL AND H. SCHNEIDER [1973], *Cyclic and diagonal products on a matrix*, Linear Algebra Appl., 7, pp. 301–335.

——— [1975], *Diagonal similarity and equivalence for matrices over groups with* 0, Czechoslovak Math. J., 25, pp. 389–403.

——— [1980], *A simple algorithm for finding a canonical form of a matrix under diagonal similarity, and applications*, abstract.

M. D. GUNZBURGER AND R. J. PLEMMONS [1979], *Energy conserving norms for the solution of hyperbolic systems of partial differential equations*, Math. Comp., 33, pp. 1–10.

T. D. HOWELL [1976], *Partitioning using* PAQ^T, in Sparse Matrix Computations, J. R. Bunch and D. J. Rose, eds., Academic Press, New York.

D. B. SAUNDERS AND H. SCHNEIDER [1978], *Flows on graphs applied to diagonal similarity and diagonal equivalence for matrices*, Discrete Math., 24, pp. 205–220.

R. SINKHORN AND P. KNOPP [1969], *Problems concerning diagonal products in nonnegative matrices*, Trans. Amer. Math. Soc., 136, pp. 67–75.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 1, March 1981

0196-5212/81/0201-0009 $01.00/0

NONNEGATIVE λ-MONOTONE MATRICES*

S. K. JAIN† AND L. E. SNYDER†

Abstract. In this paper, the structure of nonnegative $m \times n$ matrices A satisfying $AXA = A$, for some nonnegative $n \times m$ matrix X, is obtained. Several equivalent characterizations of such matrices A have been given earlier by Plemmons [Proc. Amer. Math. Soc., 39 (1973), pp. 26–32] and Berman–Plemmons [Linear and Multlinear Algebra, 2 (1974), pp. 161–172]. The structure of matrices given in this paper unifies all the previous known results on λ-monotone matrices where $1 \in \lambda$. The importance of λ-monotonicity to problems in mathematical economics, in probability and statistics, and in numerical linear algebra, is documented in a recent book by Berman and Plemmons [*Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979].

1. Introduction. Let A be an $m \times n$ real matrix. Consider the equations: (1) $AXA = A$, (2) $XAX = X$, (3) $(AX)^T = AX$, (4) $(XA)^T = XA$, and (5) $AX = XA$, where X is an $n \times m$ real matrix and T denotes the transpose. Let λ be a nonempty subset of $\{1, 2, 3, 4, 5\}$. X is called a λ-inverse of A if X satisfies equation (i) for each $i \in \lambda$. A λ-inverse of a matrix A is generally denoted by $A^{(\lambda)}$. A $\{1, 2, 3, 4\}$–inverse of A is the unique *Moore–Penrose inverse* of A and is denoted by $A^{\dagger}$. A $\{1, 2, 5\}$–inverse of A exists if and only if $m = n$ and rank A = rank A^2, i.e., index $A = 1$, and is denoted by $A^{\#}$.

A matrix $A = (a_{ij})$ is *nonnegative* if $a_{ij} \geqq 0$ for all i, j, and we denote it by $A \geqq 0$. If $a_{ij} > 0$ for all i, j, we write $A > 0$. A nonnegative matrix is called *λ-monotone* if its λ-inverse is nonnegative. If a matrix A is a direct sum of matrices S_i, then S_i's will be called *summands* of A. S_r will denote the *symmetric group on r symbols*, say $\{1, 2, \cdots, r\}$. Diag A shall denote the main diagonal of the matrix A.

Nonnegative matrices have played a significant role in numerical analysis, economics, and Markov chains. The interested reader is referred to a wealth of selected applications of nonnegative matrices to numerical analysis, probability, economics, and operations research in a recent book by Berman and Plemmons [2]. In many of the applications, one is interested in finding nonnegative solutions of the system $Ax = b$, where $A \geqq 0$ and $b \geqq 0$. If $A^{(\lambda)} \geqq 0$ exists with $1 \in \lambda$ and if the system is consistent, then $x = A^{(\lambda)}b$ provides a nonnegative solution for the system. Of course, as is well known, the existence of the nonnegative $\{1\}$–inverse is sufficient but not necessary for obtaining a nonnegative solution to the consistent system. Also, in the case where the system is inconsistent if B is a $\{1, 3\}$–inverse, then $x = Bb$ yields a least squares solution; i.e., the minimum of $\|Ax - b\|_2$ is attained for $x = Bb$.

Theorem 1 gives the structure of matrices $A \geqq 0$ having a nonnegative $\{1\}$–inverse. The representation of matrices A obtained in Theorem 1 provides a new proof for theorems of Plemmons [17], Berman–Plemmons [5], and gives immediately as special cases the theorems of Berman [3], Plemmons–Cline [18], Haynsworth–Wall [10], [11], Jain–Goel–Kwak [12], [13], [14], Lewin [16] and perhaps some others (see Corollaries 3, 4). Our main result makes use of the following:

THEOREM A [7, Theorem 2]. *If E is a nonnegative idempotent matrix of rank r, then there exists a permutation matrix P such that*

$$PEP^T = \begin{bmatrix} J & JD & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CJ & CJD & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

* Received by the editors March 31, 1980, and in revised form July 18, 1980.
† Department of Mathematics, Ohio University, Athens, Ohio 45701.

where J is a direct sum of matrices $x_i y_I^T$, x_I, $y_I > 0$ and $y_i^T x_i = 1$ and C, D are nonnegative matrices of suitable sizes.

THEOREM B [15, Lemma 2]. *Let X, Y be respectively $m \times n$, $n \times m$ nonnegative matrices such that*

$$XY = \begin{bmatrix} X_1 & & 0 \\ & \ddots & \\ 0 & & X_1 \end{bmatrix}, \qquad YX = \begin{bmatrix} Y_1 & & 0 \\ & \ddots & \\ 0 & & Y_r \end{bmatrix},$$

where X_i, Y_i are positive square matrices of order a_i, α_i respectively.

If $X = (X_{ij})$, $Y = (Y_{ij})$ are partitionings of X, Y respectively such that X_{ii}, Y_{ii} are of orders $a_i \times \alpha_i$, $\alpha_i \times a_i$ respectively, then there exists $\sigma \in S_r$ such that $X_{j\sigma(j)} \neq 0$, $Y_{\sigma(j)j} \neq 0$, $X_{jk} = 0 = Y_{kj}$, for all $k \neq \sigma(j)$.

2. Preliminary results.

LEMMA 1. *Let L, M be nonnegative matrices of orders $m \times n$, $n \times m$, respectively, such that*

$$LM = \begin{bmatrix} K_1 & K_1D_1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ C_1K_1 & C_1K_1D_1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad ML = \begin{bmatrix} K_2 & K_2D_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ C_2K_2 & C_2K_2D_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where diagonal blocks are square matrices, C_i, D_i, $i = 1, 2$ are matrices of suitable sizes, diag $K_i > 0$, *and* rank L = rank M = rank LM = rank ML. *Let $L = (L_{ij})$, $M = (M_{ij})$, $1 \leqq i, j \leqq 4$, be partitionings of L, M such that the block multiplication of L with M in either order can be performed. Then*

$$L = \begin{bmatrix} L_{11} & L_{11}Z & 0 & 0 \\ 0 & 0 & 0 & 0 \\ XL_{11} & XL_{11}Z & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad M = \begin{bmatrix} M_{11} & M_{11}Z' & 0 & 0 \\ 0 & 0 & 0 & 0 \\ X'M_{11} & X'M_{11}Z & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

for some matrices Z, X, Z', X' (not necessarily nonnegative) of suitable sizes, and

$$\text{rank } L_{11} = \text{rank } L = \text{rank } M = \text{rank } M_{11} \quad \textit{and} \quad L_{11}M_{11} = K_1, \quad M_{11}L_{11} = K_2.$$

Proof. We have

$$L_{11}M_{11} + L_{12}M_{21} + L_{13}M_{31} + L_{14}M_{41} = K_1, \tag{1}$$

$$M_{11} + L_{11} + M_{12}L_{21} + M_{13}L_{31} + M_{14}L_{41} = K_2, \tag{2}$$

$$L_{1j}M_{j3} = 0, \quad 1 \leqq j \leqq 4, \tag{3}$$

$$M_{1j}L_{j3} = 0, \quad 1 \leqq j \leqq 4, \tag{3}'$$

$$L_{1j}M_{j4} = 0, \quad 1 \leqq j \leqq 4, \tag{4}$$

$$M_{1j}L_{j4} = 0, \quad 1 \leqq j \leqq 4, \tag{4}'$$

$$L_{2j}M_{jk} = 0, \quad 1 \leqq j, k \leqq 4, \tag{5}$$

$$M_{2j}L_{jk} = 0, \quad 1 \leqq j, k \leqq 4, \tag{5}'$$

$$L_{3j}M_{j3} = 0, \quad 1 \leqq j \leqq 4, \tag{6}$$

$$M_{3j}L_{j3} = 0, \quad 1 \leqq j \leqq 4, \tag{6}'$$

(7) $$L_{3j}M_{j4} = 0, \qquad 1 \leqq j \leqq 4,$$

(7)′ $$M_{3j}L_{j4} = 0, \qquad 1 \leqq j \leqq 4,$$

(8) $$L_{4j}M_{jk} = 0, \qquad 1 \leqq j, k \leqq 4,$$

(8)′ $$M_{4j}L_{jk} = 0, \qquad 1 \leqq j, k \leqq 4.$$

Premultiply (1) by M_{21} and use (5)′ to obtain $0 = M_{21}K_1$. Since diag $K_1 > 0$, we get $M_{21} = 0$. Postmultiply (1) by L_{13} and use (3)′, (6)′, (8)′ to get $L_{13} = 0$. Similarly, we get $M_{41} = 0 = L_{14}$.

Similar computations yield

$$L_{21} = 0 = M_{13} = L_{41} = M_{14}.$$

Thus, $L_{11}M_{11} = K_1$, $M_{11}L_{11} = K_2$.

It follows easily that

$$\text{rank } L = \text{rank } L_{11} \equiv \text{rank } K_1.$$

Therefore, for all $i = 2, 3, 4$ there exists a matrix X_i of suitable size such that

(9) $$(L_{i1} \quad L_{i2} \quad L_{i3} \quad L_{i4}) = X_i(L_{11} \quad L_{12} \quad 0 \quad 0).$$

Thus, $L_{i3} = 0 = L_{i4}$, $i = 2, 3, 4$.

Also, rank L = rank L_{11} implies that there exists a matrix Z of suitable size such that

(10) $$\begin{bmatrix} L_{12} \\ L_{22} \\ L_{32} \\ L_{42} \end{bmatrix} = \begin{bmatrix} L_{11} \\ 0 \\ L_{31} \\ 0 \end{bmatrix} Z.$$

But then, $L_{22} = 0 = L_{42}$. Hence,

$$L = \begin{bmatrix} L_{11} & L_{12} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ L_{31} & L_{32} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

From (10) above it follows that $L_{12} = L_{11}Z$, and letting $X_3 = X$ in (9) yields $L_{31} = XL_{11}$ and $L_{32} = XL_{12} = XL_{11}Z$. Similarly,

$$M = \begin{bmatrix} M_{11} & M_{12} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ M_{31} & M_{32} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where $M_{12} = M_{11}Z'$, $M_{31} = X'M_{11}$, $M_{32} = X'M_{11}Z'$ for some matrices Z' and X'. This completes the proof.

Lemma 2, which is essentially Theorem B, gives the nature of the submatrices L_{11}, M_{11} of the matrices L, M respectively appearing in Lemma 1.

LEMMA 2. *Let X, Y be $m \times n$, $n \times m$ matrices each of rank r such that*

$$XY = \begin{bmatrix} a_1 b_1^T & & 0 \\ & \ddots & \\ 0 & & a_r b_r^T \end{bmatrix}, \qquad YX = \begin{bmatrix} c_1 d_1^T & & 0 \\ & \ddots & \\ 0 & & c_r d_r^T \end{bmatrix},$$

where $a_i, b_i, c_i, d_i > 0$, diagonal blocks are square matrices and a_i, a_j (and c_i, c_j), $i \neq j$, are not necessarily of the same sizes. Then:

(1) *There exist permutation matrices P, Q of orders m, n respectively, such that PXQ^T is a direct sum of matrices of types* (I), (II) (*not necessarily both*):

(I) βxy^T, $\beta > 0$, x, y, *are positive unit vectors.*

(II)
$$\begin{bmatrix} 0 & \beta_{12}x_1y_2^T & 0 & \cdots & 0 \\ 0 & 0 & \beta_{23}x_2y_3^T & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \beta_{d-1,d}x_{d-1}y_d^T \\ \beta_{d1}x_dy_1^T & 0 & 0 & \cdots & 0 \end{bmatrix}$$

with all β's > 0; x_i, y_j are positive unit vectors, not necessarily of the same size.

(2) *X has a nonnegative $\{1, 2\}$-inverse.*

Similar results hold for Y.

Furthermore, $P = Q$ if $m = n$.

Proof. The proof, although straightforward, is rather technical and is omitted.

Note. If S is a summand of type (I) then one can verify that

$$\beta^{-1}yx^T$$

is a $\{1, 2\}$-inverse of S, whereas if S is a summand of type (II) then

$$\begin{bmatrix} 0 & 0 & \cdots & 0 & \beta_{d1}^{-1}y_1x_d^T \\ \beta_{12}^{-1}y_2x_1^T & 0 & \cdots & 0 & 0 \\ 0 & \beta_{23}^{-1}y_3x_2^T & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \beta_{d-1,d}^{-1}y_dx_{d-1}^T & 0 \end{bmatrix}$$

is a $\{1, 2\}$-inverse of S.

Henceforth, by matrices of types (I) or (II) we shall mean the matrices of types (I) or (II) described in Lemma 2.

3. Main results.

THEOREM 1. *Let A be a nonnegative $m \times n$ matrix. Then A has a nonnegative $\{1\}$-inverse X if and only if for some permutation matrices P, Q of orders m, n, respectively,*

$$PAQ^T = \begin{bmatrix} J & JD & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CJ & CJD & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where J is a direct sum of matrices of types (I) *and* (II) (*not necessarily both*), *and C, D are nonnegative matrices of suitable sizes.*

Proof. We have $AXA = A$. This gives $\operatorname{rank} A = \operatorname{rank} AX = \operatorname{rank} XA = \operatorname{rank} XAX = r$, say. Further, since AX and XA are nonnegative idempotents we have, by Flor [7],

$$P_1AXP_1^T = \begin{bmatrix} K_1 & K_1D_1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ C_1K_1 & C_1K_1D_1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{11}$$

and

$$Q_1XAQ_1^T = \begin{bmatrix} K_2 & K_2D_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ C_2K_2 & C_2K_2D_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{12}$$

where P_1, Q_1 are permutation matrices of orders m, n respectively, C_1, D_1, C_2, D_2 are nonnegative matrices of suitable sizes, and each K_i, $i = 1, 2$, is a square matrix of rank r which is a direct sum of matrices of the form $x_iy_i^T$ and x_i, y_i are positive unit vectors with $y_i^Tx_i = 1$. Set $L = P_1AQ_1^T$, $M = Q_1(XAX)P_1^T$. Then $LM = P_1AXP_1^T$ and $ML = Q_1XAQ^T$. Also, since rank A = rank AX = rank XA = rank XAX, we have

$$\text{rank } L = \text{rank } LM = \text{rank } ML = \text{rank } M.$$

Thus, by Lemma 1, L, M are of the forms

$$L = \begin{bmatrix} L_{11} & L_{11}Z & 0 & 0 \\ 0 & 0 & 0 & 0 \\ XL_{11} & XL_{11}Z & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad M = \begin{bmatrix} M_{11} & M_{11}Z' & 0 & 0 \\ 0 & 0 & 0 & 0 \\ X'M_{11} & X'M_{11}Z' & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where

$$L_{11}M_{11} = K_1, M_{11}L_1 = K_2. \tag{13}$$

Then, by Lemma 2, there exist permutation matrices P_2, Q_2 of suitable orders such that $P_2L_{11}Q_2^T$ is a direct sum of matrices of the types (I), (II) stated in the theorem. Also, by Lemma 2, L_{11} possesses a nonnegative $\{1\}$-inverse $L_{11}^{(1)}$. Thus, if we set $L_{11}^{(1)}L_{11}Z = D'$, $XL_{11}L_{11}^{(1)} = C'$, then D', $C' \geqq 0$ and $L_{11}Z = L_{11}D'$, $XL_{11} = C'L_{11}$, $XL_{11}Z = C'L_{11}D'$.

Finally, let us set

$$P = \begin{bmatrix} P_2 & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix} P_1, \qquad Q = \begin{bmatrix} Q_2 & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix} Q_1, \tag{14}$$

where the I's are identity matrices of suitable orders such that P, Q are of orders m, n respectively, and the partitionings of P, Q given above are such that the block multiplication of $PAQ^T = PP_1^TLQ_1Q^T$ can be performed. Then

$$PAQ^T = \begin{bmatrix} J & JD & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CJ & CJD & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where $D = Q_2D'$, $C = C'P_1^T$ are nonnegative matrices of suitable sizes and $J = P_2L_{11}Q_2^T$ is a direct sum of matrices of types (I) and (II) (not necessarily both) as obtained above, completing the "only if" part. To prove the "if part" we observe (see note following Lemma 2) that if S denotes a summand of J, then it has a nonnegative

{1}-inverse, $S^{(1)}$. Consequently, J has a nonnegative {1}-inverse $J^{(1)}$. Also, since

$$Q^T \begin{bmatrix} J^{(1)} & J^{(1)}D & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CJ^{(1)} & CJ^{(1)}D & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} P$$

is a nonnegative {1}-inverse of A, the converse follows.

COROLLARY 1. *Let* $\lambda = \{1, 5\}$. *Then a nonnegative square matrix* A *is* λ*-monotone if and only if there exists a permutation matrix* P *such that*

$$PAP^T = \begin{bmatrix} J & JD & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CJ & CJD & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where C, D *are nonnegative matrices of suitable size, diagonal blocks are square matrices and* J *is a direct sum of matrices of the following types (not necessarily both):*

(I)$_*$ βxy^T, $\beta > 0$, x, y *are positive unit vectors of the same size and* $y^T x = 1$.

(II)$_*$

$$\begin{bmatrix} 0 & \beta_{12}x_1y_2^T & 0 & \cdots & 0 \\ 0 & 0 & \beta_{23}x_2y_3^T & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \beta_{d-1,d}x_{d-1}y_d^T \\ \beta_{d1}x_dy_1^T & 0 & 0 & \cdots & 0 \end{bmatrix}$$

with $\beta_{ij} > 0$; x_i, y_i *are positive unit vectors,* x_i, y_i *are of the same size,* x_i, y_j, $i \neq j$ *are not necessarily of the same size and* $y_i^T x_i = 1$.

Proof. Let X be a nonnegative {1, 5}-inverse of A. Then $AXA = A$, $AX = XA$. Clearly A, X are square matrices of the same order. Thus, in the equations (11) and (12) in the proof of the theorem, we have $P_1 = Q_1$, and L, M (as well as L_{11}, M_{11}) are of the same order. Then by the last statement in Lemma 2, and equation (13) in the theorem, we get $P_2 = Q_2$. Hence, by (14) in the theorem, $P = Q$. Thus,

$$PAP^T = \begin{bmatrix} J & JD & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CJ & CJD & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where C, D are nonnegative matrices of suitable sizes and J is a square matrix. That J is a direct sum of matrices of the types stated in the corollary is obvious, completing the "only if" part of the corollary. The "if part" follows as in the proof of the theorem.

COROLLARY 2. *The class of nonnegative* {1}*-monotone matrices coincides with the class of nonnegative* {1, 2}*-monotone matrices.*

Proof. Let $\mathscr{C}_1$ denote the class of nonnegative {1}-monotone matrices and C_2 denote the class of nonnegative {1, 2}-monotone matrices.

Let $A \in \mathscr{C}_1$. Then, by Theorem 1, there exist permutation matrices P, Q of suitable orders such that

$$PAQ^T = \begin{bmatrix} J & JD & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CJ & CJD & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where C, $D \geqq 0$ and J is a direct sum of matrices of types (I) and (II). Now, if S is a summand of J, then as noted before, S has a nonnegative $\{1, 2\}$-inverse $S^{(1,2)}$. Hence, $J^{(1,2)} \geqq 0$.

Since

$$A^{(1,2)} = P^T \begin{bmatrix} J^{(1,2)} & J^{(1,2)} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CJ^{(1,2)} & CJ^{(1,2)} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} Q,$$

it follows that $A^{(1,2)} \geqq 0$; i.e., $A \in \mathscr{C}_2$. Hence, $\mathscr{C}_1 = \mathscr{C}_2$.

COROLLARY 3. *Let A be a nonnegative matrix and let $A^{(1)} = p(A) \geqq 0$, where $p(A) = \sum_{i=1}^{k} \alpha_i A^{m_i}$, $\alpha_i \neq 0$, $m_i \geqq 0$. Then there exists a permutation matrix P such that*

$$PAP^T = \begin{bmatrix} J & JD & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CJ & CJD & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where C, D are nonnegative matrices of appropriate sizes and J is a direct sum of matrices of the following types (not necessarily both):

(I)$_{**}$ *βxy^T, where x and y are positive unit vectors with $y^T x = 1$ and β is a positive root of*

(a)
$$\sum_{i=1}^{k} \alpha_i t^{m_i+1} = 1.$$

(II)$_{**}$
$$\begin{bmatrix} 0 & \beta_{12}x_1y_2^T & 0 & \cdots & 0 \\ 0 & 0 & \beta_{23}x_2y_3^T & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & & \beta_{d-1,d}x_{d-1}y_d^T \\ \beta_{d1}x_dy_1^T & 0 & 0 & \cdots & 0 \end{bmatrix},$$

where x_i, y_i are positive unit vectors of the same order with $y_i^T x_i = 1$; x_i and x_j, $i \neq j$, are not necessarily of the same order. $\beta_{12}, \cdots, \beta_{d1}$ are arbitrary positive numbers with $d > 1$ and $d | m_i + 1$ for some m_i such that the product $\beta_{12}\beta_{23} \cdots \beta_{d1}$ is a common root of the following system of at most d equations in t:

(b)
$$\sum_{d \in \Lambda_0} \alpha_i t^{(m_i+1)/d} = 1,$$

$$\sum_{d \in \Lambda_k} \alpha_i t^{(m_i+1-k)/d} = 0, \qquad k \in \{1, 2, \cdots, d-1\},$$

where

$$\Lambda_k = \{d : d | m_i + 1 - k,\ d \neq 1\}, \qquad k = 0, 1, \cdots, d-1,$$

with the understanding that if some $\Lambda_k = \varnothing$ *then the corresponding equation is absent.*

Conversely, suppose we have, for some permutation matrix P,

$$PAP^T = \begin{bmatrix} J & JD & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CJ & CJD & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where C, D are arbitrary nonnegative matrices of appropriate sizes and J is a direct sum of matrices of the following types (not necessarily both):

(I′) βxy^T, $\beta > 0$, x, y *are positive vectors with* $y^T x = 1$.

(II′)
$$\begin{bmatrix} 0 & \beta_{12}x_1y_2^T & 0 & 0 & \cdots & 0 \\ 0 & 0 & \beta_{23}x_2y_3^T & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & \beta_{d-1,d}x_{d-1}y_d^T \\ \beta_{d1}x_dy_1^T & 0 & 0 & \cdots & 0 & 0 \end{bmatrix},$$

where $\beta_{ij} > 0$, x_i *and* y_i *are positive vectors with* $y_i^T x_i = 1$. *Then* $A^{(1,2)} \geqq 0$ *and is equal to some polynomial in A with scalar coefficients.*

Proof. By Corollary 1, there exists a permutation matrix P such that

$$PAP^T = \begin{bmatrix} J & JD & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ CJ & CJD & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

where J is a direct sum of matrices of types $(\mathrm{I})_*$ and $(\mathrm{II})_*$ (not necessarily both).

Since $p(A) = \sum_{i=1}^{k} \alpha_i A^{m_i}$ is a $\{1\}$-inverse of A,

$$\alpha_1 A^{m_1+2} + \cdots + \alpha_k A^{m_k+2} = A. \tag{15}$$

Also, it is straightforward to verify that if $f(A)$ is any polynomial in A with scalar coefficients, then

$$Pf(A)P^T = \begin{bmatrix} f(J) & f(J)D & 0 & 0 \\ 0 & 0 & 0 & 0 \\ Cf(J) & Cf(J)D & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Thus, (1) implies

$$\alpha_i J^{m_1+2} + \cdots + \alpha_k J^{m_k+2} = J. \tag{16}$$

Clearly, all summands S of J will also satisfy (2); i.e.,

$$\alpha_1 S^{m_1+2} + \cdots + \alpha_k S^{m_k+2} = S. \tag{17}$$

Then it is a direct verification that if S is a summand of type $(\mathrm{I})_*$ then β must satisfy the equation (a), and if S is a summand of type $(\mathrm{II})_*$ then $\beta_{12}\beta_{23}\cdots\beta_{d1}$ must satisfy the system of equations (b). Hence, J is a direct sum of matrices of the form $(\mathrm{I})_{**}$ and $(\mathrm{II})_{**}$ as desired (for details see [13, Theorem 2]).

Remark 1. The above corollary gives in particular, the earlier known results of Harary-Minc [9], Berman [3], Lewin [16], Jain–Goel–Kwak [12], [13], [14], Haynsworth–Wall [10], [11].

The following theorem giving equivalent characterizations of {1}-monotone matrices was first proved by Berman–Plemmons and also earlier by Plemmons for square matrices. We show that those characterizations may be obtained as a direct consequence of the structure theorem.

THEOREM 2. [5, Theorem 4]. *For an $m \times n$ nonnegative matrix A of rank r the following are equivalent*:

(a) *A is {1}-monotone.*

(b) *There exists a {1}-inverse of the form $D_1 A^T D_2$, where D_1, D_2 are nonnegative diagonal matrices.*

(c) *A has a monomial submatrix of rank r.*

(d) *A has a nonnegative rank factorization FG where F, G have monomial submatrices of rank r.*

Proof. (a)$\Rightarrow$(b). By Theorem 1,

$$PAQ^T = \begin{bmatrix} J & JD & 0 & 0 \\ 0 & 0 & 0 & 0 \\ CJ & CJD & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

for some permutation matrices P, Q, and J is a direct sum of matrices of types (I) and (II) (not necessarily both).

If S is a summand of J, then it can be verified that

$$S^{(1)} = \begin{cases} \beta^{-2} S^t & \text{if } S \text{ is of type (I),} \\ \operatorname{diag}(\beta_{d1}^{-2}, \beta_{12}^{-2}, \cdots, \beta_{d-1,d}^{-2}) S^T & \text{if } S \text{ is of type (II).} \end{cases}$$

Thus, $J^{(1)} = ZJ^T$, where Z is a diagonal matrix. Further, direct verification yields that

$$A^{(1)} = Q^T \begin{bmatrix} J^{(1)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} P$$

$$= Q^T \begin{bmatrix} ZJ^T & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} P$$

$$= Q^T \begin{bmatrix} Z & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} QQ^T \begin{bmatrix} J^T & 0 & J^T C^T & 0 \\ D^T J^T & 0 & D^T J^T C^T & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} P^T P \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} P^T$$

$$= D_1 A^T D_2,$$

where

$$D_1 = Q^T \begin{bmatrix} Z & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} Q,$$

and

$$D_2 = P\begin{bmatrix} I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} P^T$$

are nonnegative diagonal matrices. This proves (a)⇒(b). (b)⇒(a) is obvious.

(a)⇔(c). (a)⇒(c) follows at once from the representation of {1}–monotone matrices. (c)⇒(a) is easy and is left to the reader. (Note (c) implies that there are permutation matrices, P, Q such that $PAQ^T = \left[\begin{smallmatrix} M & X \\ Y & Z \end{smallmatrix}\right]$, where rank A = rank M, and $M^{-1} \geqq 0$.)

(a)⇒(d). We appeal to Theorem 1 here again. First we note that if $J = FG$ is a nonnegative rank factorization of J, then it "lifts" to a nonnegative rank factorization

$$A = P^T\begin{bmatrix} F \\ 0 \\ CF \\ 0 \end{bmatrix}(G \quad GD \quad 0 \quad 0)Q$$

of A. Furthermore, it is clear that if S is a summand of J of type (I) or of type (II), then S has a nonnegative rank factorization $S = FG$, such that F and G contain monomials of rank r. This proves (a)⇒(d) (for details see [14, Theorems 3, 4]).

(d)⇒(a). This is well known and is left to the reader.

Remark 2. Using the representation of matrices obtained in Theorem A, the other results of Berman–Plemmons in [5] for λ-monotone matrices where $1 \in \lambda$ and the theorems of Plemmons–Cline [17] for the nonnegativity of the Moore–Penrose inverse can be similarly obtained.

Summary. This paper unifies all previously known results on nonnegative λ-monotone matrices where $1 \in \lambda$. The main theorem gives the structure of nonnegative matrices having a nonnegative {1}-inverse as a direct sum of certain "well-defined blocks" (of types (I) and (II)). One of the problems which has led to some of the interest in λ-monotone matrices is the problem of obtaining nonnegative solutions of a linear system $Ax = b$. As is well known, if X is a {1}-inverse (or {1, 3}-inverse) of A, then $x = Xb$ is a solution of $Ax = b$ if the system is consistent (or a best approximate solution), and so obviously Xb is nonnegative whenever X and b are both nonnegative. Of course, there is still much that remains to be done in order to characterize systems having nonnegative solutions. A recent paper by S. Friedland and H. Schneider [8] addresses this question.

REFERENCES

[1] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications*, Wiley, New York, 1974.

[2] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[3] A. Berman, *Nonnegative matrices which are equal to their generalized inverse*, Linear Algebra Appl., 9 (1974), pp. 261–265.

[4] A. Berman and R. J. Plemmons, *Monotonicity and the generalized inverse*, SIAM J. Appl. Math., 22 (1972), pp. 155–161.

[5] ———, *Inverses of nonnegative matrices*, Linear and Multilinear Algebra, 2 (1974), pp. 161–172.

[6] R. E. CLINE, *Inverses of rank invariant powers of a matrix*, SIAM J. Numer. Anal., 5 (1968), pp. 182–197.
[7] P. Flor, *On groups of nonnegative matrices*, Compositio Math., 21 (1969), pp. 376–382.
[8] S. FRIEDLAND AND H. SCHNEIDER, *The growth of powers of a nonnegative matrix*, this Journal, 1 (1980), pp. 185–200.
[9] F. HARARY AND H. MINC, *Which nonnegative matrices are self-inverse*, Math. Mag., 49 (1976), pp. 91–92.
[10] E. HAYNSWORTH AND J. R. WALL, *Group inverses of certain nonnegative matrices*, Linear Algebra Appl., 25 (1979), pp. 271–288.
[11] E. HAYNSWORTH AND J. R. WALL, *Group inverses of certain positive operators*, preprint.
[12] S. K. JAIN, V. K. GOEL AND EDWARD K. KWAK, *Nonnegative mth roots of nonnegative 0-symmetric idempotent matrices*, Linear Algebra Appl., 23 (1979), pp. 37–51.
[13] ———, *Nonnegative matrices having some nonnegative Moore-Penrose and group inverses*, Linear and Multilinear Algebra, 7 (1979), pp. 59–72.
[14] S. K. JAIN, EDWARD K. KWAK AND V. K. GOEL, *Decomposition of nonnegative group-monotone matrices*, Trans. Amer. Math. Soc., 257 (1980), pp. 371–385.
[15] S. K. JAIN, *Nonnegative matrices having nonnegative W-weighted group inverses*, submitted.
[16] M. LEWIN, *Nonnegative matrices generating a finite cycle group*, Linear and Multilinear Algebra, 5 (1977), pp. 91–94.
[17] R. J. PLEMMONS, *Regular nonnegative matrices*, Proc. Amer. Math. Soc., 39 (1973), pp. 26–32.
[18] R. J. PLEMMONS AND R. E. CLINE, *The generalized inverse of a nonnegative matrix*, Proc. Amer. Math. Soc., 31 (1972), pp. 46–50.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 1, March 1981

0196-5212/81/0201-0010 $01.00/0

COMPUTING THE MINIMUM FILL-IN IS NP-COMPLETE

MIHALIS YANNAKAKIS†

Abstract. We show that the following problem is NP-complete. Given a graph, find the minimum number of edges (fill-in) whose addition makes the graph chordal. This problem arises in the solution of sparse symmetric positive definite systems of linear equations by Gaussian elimination.

1. Introduction and terminology. A graph is a pair $G = (N, E)$, where N is a finite set of *nodes* and E, a set of unordered pairs (u, v) of distinct nodes, is a set of *edges*. Two nodes u and v are *adjacent* if $(u, v) \in E$. The *neighborhood* $\Gamma(v)$ of a node v is the set of nodes that are adjacent to v. The *degree* $d(v)$ of v is the number of nodes adjacent to v. A graph is a *clique* if every two nodes are adjacent. A set of nodes is *independent* if no two of them are adjacent.

If $S \subseteq N$ is a subset of nodes, the *subgraph* of G *induced* by S, denoted as $\langle S \rangle$, is the graph (S, E_S), where $E_S = \{(u, v) \in E | u, v \in S\}$. The graph $G - S$, formed by deleting a subset $S \subseteq N$ of nodes from G, is $\langle N - S \rangle$. A graph $G = (N, E)$ is *bipartite* if N can be partitioned into two sets P, Q of independent nodes; we will write the bipartite graph as (P, Q, E). The bipartite graph (P, Q, E) is a *chain graph* if the neighborhoods of the nodes in P form a chain; i.e., there is a bijection $\pi: \{1, 2, \cdots, |P|\} \leftrightarrow P$ (an *ordering* of P) such that $\Gamma(\pi(1)) \supseteq \Gamma(\pi(2)) \supseteq \cdots \supseteq \Gamma(\pi(|P|))$. It is easy to see [Y] that then the neighborhoods of the nodes in Q form also a chain, and thus the definition is unambiguous.

A graph is *chordal* (or *triangulated*) if every cycle of length $\geqq 4$ has a *chord*, i.e., an edge connecting two nonconsecutive nodes of the cycle. Chordal graphs are important in connection with the solution of sparse symmetric positive definite systems of linear equations by Gaussian elimination [R]. From the symmetric $n \times n$ matrix $M = (m_{ij})$ of coefficients of such a system we can construct a graph $G = (N, E)$ with n nodes, where node v_i corresponds to the ith row and column of M and $(v_i, v_j) \in E$ iff $m_{ij} \neq 0$. The *elimination* of node v_i from G is performed by (1) adding edges so that $\Gamma(v_i)$ becomes a clique, and (2) deleting v_i from the augmented graph. The added edges correspond to the new nonzero elements that are created when we eliminate the ith variable, assuming no lucky cancellations. (See [R] for a detailed exposition of this graph-theoretic modeling.) If π is an ordering of N, the *fill-in* $F(\pi)$ produced by π is the set of new edges that are added when we eliminate $\pi(1)$ from G, then eliminate $\pi(2)$ from the resulting graph, $\pi(3)$ from the new graph, etc. The ordering π is a *perfect elimination ordering* if $F(\pi) = \varnothing$. Chordal graphs come into the picture because of the following two properties [R]. (1) A graph has a perfect elimination ordering if and only if it is chordal. Thus, "chordal" is a *hereditary* property (i.e., deleting nodes from a chordal graph does not violate the property), and every chordal graph has a node v such that $\langle \Gamma(v) \rangle$ is a clique; v is called a *simplicial* node. (2) If π is an elimination ordering of a graph $G = (N, E)$, then the augmented graph $G_\pi = (N, E \cup F(\pi))$ is chordal: π is a perfect elimination ordering of G_π.

In this paper we examine the problem of finding an elimination ordering which produces a minimum fill-in, or equivalently, finding the minimum set of edges whose addition renders the graph chordal. We shall show that this problem is NP-complete.

* Received by the editors June 26, 1980.
† Bell Laboratories, Murray Hill, New Jersey 07974.

(For an exposition of NP-completeness see [GJ].) The NP-completeness of the minimum fill-in problem was conjectured in [RTL] and [RT], but a proof had not been found, and it is one of the open problems in [GJ]. The version of the problem on directed graphs was shown to be NP-complete in [RT].

2. The reduction. We will make use of chain graphs. Two edges (u, v), (x, y) are said to be *independent* in a graph G if the nodes u, v, x, y are distinct and the subgraph of G induced by them consists of exactly these two edges. The following lemma from [Y] is easy to prove.

LEMMA 1. *A bipartite graph is a chain graph if and only if it does not contain a pair of independent edges.*

Let $G=(P, Q, E)$ be a bipartite graph. From G we construct another graph $C(G)=(N, E')$ by making P and Q cliques; i.e., $E'=E\cup\{(u, v)|u, v\in P\}\times \cup\{(u, v)|u, v\in Q\}$.

LEMMA 2. *Let G be a bipartite graph. $C(G)$ is chordal if and only if G is a chain graph.*

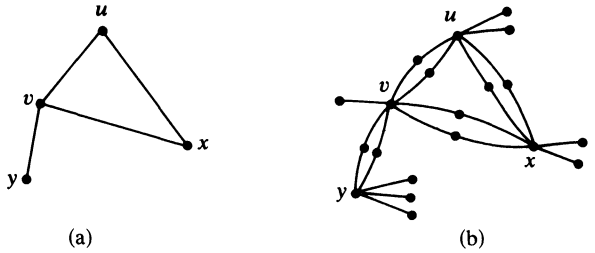
Proof (*only if*). Suppose that G is not a chain graph. Then it has two independent edges (u, v) and (x, y) by Lemma 1. Suppose without loss of generality that $u, x\in P$ and $v, y\in Q$. Then these two edges together with (u, x) and (v, y) form a chordless cycle of length 4 in $C(G)$.

(*if*). Suppose that G is a chain graph, and let π be an ordering of P such that $\Gamma(\pi(1))\supseteq\Gamma(\pi(2))\supseteq\cdots\supseteq\Gamma(\pi(p))$, where $p=|P|$. Since the property of being a chain graph is hereditary, it suffices to show that $C(G)$ has a simplicial node. The neighborhood of $\pi(p)$ in $C(G)$ is $\Gamma'(\pi(p))=\Gamma(\pi(p))\cup[P-\pi(p)]$. In $C(G)$ the subgraphs $\langle P-\pi(p)\rangle$ and $\langle\Gamma(\pi(p))\rangle$ are cliques, the latter because $\Gamma(\pi(p))\subseteq Q$ and $\langle Q\rangle$ is a clique. Also, since $\Gamma(\pi(p))\subseteq\Gamma(v)$ for every $v\in P$, all nodes of P are adjacent to all nodes of $\Gamma(\pi(p))$. Therefore $\langle\Gamma'(\pi(p))\rangle$ is a clique, and $\pi(p)$ is a simplicial node of $C(G)$. □

LEMMA 3. *It is* NP*-complete to find the minimum number of edges whose addition to a bipartite graph $G=(P, Q, E)$ gives a chain graph.*

Proof. The reduction is from the Optimal Linear Arrangement Problem. A linear arrangement of a graph $G=(N, E)$ is an ordering π of N. For an edge $e=(u, v)$ of G, let $\delta(e, \pi)=|\pi^{-1}(u)-\pi^{-1}(v)|$. The cost $c(\pi)$ of the linear arrangement π is $c(\pi)=\sum_{e\in E}\delta(e, \pi)$. The optimal linear arrangement problem is to decide, given a graph G and an integer k, whether there exists a linear arrangement π of G with cost $c(\pi)\leqq k$. This problem was shown to be NP-complete in [GJS].

Let $(G=(N, E); k)$ be an instance of the optimal linear arrangement problem. We construct a bipartite graph $G'=(P, Q, E')$ as follows. P has one node for every node of G (i.e., $P=N$); Q has two nodes e_1, e_2 for every edge e of G, and a set $R(v)$ of $n-d(v)$ nodes for every node v of N, where $n=|N|$ and $d(v)$ is the degree of v in G. If $e=(u, v)$ is an edge of G, then the nodes e_1, e_2 that correspond to e are adjacent to u and v. The nodes in $R(v)$ are adjacent to v. In Fig. 1 we show an example of this construction.



(a) *A graph G.* (b) *The graph G'.*

FIG. 1

Let $l(G)$ be the minimum cost of a linear arrangement of G, and $h(G')$ the minimum number of edges whose addition to G' gives a chain graph. We claim that

$$h(G') = l(G) + \frac{n^2(n-1)}{2} - 2m, \tag{1}$$

where n, m are respectively the numbers of nodes and edges of G. Thus, $l(G) \leqq k$ iff $h(G') \leqq k + (n^2(n-1)/2) - 2m$.

First observe that an ordering π of N specifies uniquely a minimal set $H(\pi)$ of edges whose addition makes G' a chain graph with the neighborhoods of the nodes in $P(=N)$ ordered according to π. For every node x in Q, let $\sigma(x) = \max\{i | (x, \pi(i)) \in E'\}$. Then $H(\pi) = \{(x, \pi(j)) | x \in Q, j < \sigma(x)\} - E'$. Conversely, suppose that F is a set of edges such that $G'(F) = (P, Q, E' \cup F)$ is a chain graph and let π be an ordering of the nodes in P according to their neighborhoods in $G'(F)$. It is easy to see that $F \supseteq H(\pi)$, and therefore if F is a minimal augmentation then $F = H(\pi)$. Let $h(\pi) = |H(\pi)|$. In order to show (1), it suffices thus to show that for every ordering π of N, $h(\pi) = c(\pi) + (n^2(n-1)/2) - 2m$, where $c(\pi)$ is the cost of the linear arrangement π of G.

Let π be an ordering of N. For every $v \in N$ and $x \in R(v)$, $H(\pi)$ contains $\pi^{-1}(v) - 1$ edges incident to x. Let $e = (u, v)$ be an edge of G, and suppose without loss of generality that $\pi^{-1}(u) < \pi^{-1}(v)$. The number of edges of $H(\pi)$ incident to each of the two nodes e_1, e_2 that correspond to e is $\pi^{-1}(v) - 2 = \pi^{-1}(u) + [\pi^{-1}(v) - \pi^{-1}(u)] - 2 = \pi^{-1}(u) + \delta(e, \pi) - 2$; thus, the number of edges of $H(\pi)$ incident to e_1 and e_2 is $\pi^{-1}(v) + \pi^{-1}(u) + \delta(e, \pi) - 4$. Consequently,

$$\begin{aligned} h(\pi) &= \sum_{v \in N} \sum_{x \in R(v)} [\pi^{-1}(v) - 1] + \sum_{e=(u,v) \in E} [\pi^{-1}(v) + \pi^{-1}(u) + \delta(e, \pi) - 4] \\ &= \sum_{v \in N} (n - d(v))(\pi^{-1}(v) - 1) + \sum_{v \in N} d(v)\pi^{-1}(v) + \sum_{e \in E} \delta(e, \pi) - 4m \\ &= \sum_{v \in N} n[\pi^{-1}(v) - 1] + \sum_{v \in N} d(v) + c(\pi) - 4m \\ &= c(\pi) + \frac{n^2(n-1)}{2} - 2m, \end{aligned}$$

since $\sum_{v \in N} d(v) = 2m$, and

$$\sum_{v \in N} [\pi^{-1}(v) - 1] = 0 + 1 + 2 + \cdots + (n-1) = \frac{n(n-1)}{2}. \qquad \square$$

THEOREM 1. *The minimum fill-in problem is* NP*-complete.*

Proof. Follows from Lemmas 2 and 3. $\square$

REFERENCES

[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.

[R] D. J. ROSE, *A graph- theoretic study of the numerical solution of sparse positive definite systems of linear equations*, in Graph Theory and Computing, R. Read, ed., Academic Press, New York, 1973, pp. 183–217.

[RT] D. J. ROSE AND R. E. TARJAN, *Algorithmic aspects of vertex elimination on directed graphs*, SIAM J. Appl. Math., 34 (1978), pp. 176–197.

[RTL] D. J. ROSE, R. E. TARJAN AND G. S. LUEKER, *Algorithmic aspects of vertex elimination on graphs*, SIAM J. Comput., 5 (1976), pp. 266–283.

[Y] M. YANNAKAKIS, *Node-deletion problems on bipartite graphs*, SIAM J. Comput., 10 (1981).

SIAM. J. ALG. DISC. METH.
Vol. 2, No. 2, June 1981

0196-5212/81/1202-0001 $01.00/0

A BOUNDARY PROBLEM FOR GROUP TESTING*

M. C. HU,† F. K. HWANG‡ AND JU KWEI WANG§

Abstract. A previous result [F. K. Hwang, Tamkang J. Math., 2 (1971), pp. 39–44] showed that a minimax group testing algorithm to find d defectives in n items is to test each item individually for $d \geqq 0.5n$. In this paper we improve this result by proving that individual testing is minimax for $d \geqq 0.4n$. We also conjecture that the same is true for $d \geqq \frac{1}{3}n$. On the other hand, we prove that individual testing is not minimax for $d < \frac{1}{3}n$.

1. Introduction. Suppose that a population is known to consist of n items including d defectives. A test is available to verify whether a given item is good or defective. Furthermore, suppose that a group test is available which tests a subset of items simultaneously, with two possible outcomes: a *pure outcome* indicates that all items in the subset are good, and a *contaminated outcome* indicates that at least one item in the subset is defective. An interesting question is to determine for which values of n and d the use of group tests reduces the number of tests needed to identify the d defectives. To be more specific, let $M_T(n, d)$ denote the maximum number of tests required by the algorithm T to identify the defectives in a population with given parameters d and n, where the maximum is taken over all possible combinations of the d defectives among the n items. Let I denote the algorithm which tests each item individually. Then clearly $M_I(n, d) = n - 1$ for $n > d > 0$, since the nature of the last item can always be deduced without testing but no deduction is possible with fewer than $n - 1$ items being tested in the worst case (the last two items consist of one good and one defective). Define

$$M(n, d) = \min_T M_T(n, d).$$

The question is, then, for what values of n and d is it the case that

$$M(n, d) = n - 1.$$

It was proved in [1] that

$$M(n, d) = n - 1 \quad \text{for } 2d + 1 \geqq n.$$

In this paper we improve the above result by showing that

$$M(n, d) = n - 1 \quad \text{for } 5d + 1 \geqq 2n.$$

We also conjecture that

$$M(n, d) = n - 1 \quad \text{for } 3d \geqq n.$$

We show that this is the sharpest result possible by proving that

$$M(n, d) < n - 1 \quad \text{for } 3d < n.$$

2. Some preliminary remarks. A *binary tree* is a rooted tree where each node except the root has one inlink (the root has none), and each node has either zero or two outlinks. Nodes with zero outlinks are called *terminal nodes* and nodes with two outlinks are called *internal nodes*. The *path* for a node v is the alternate sequence of

* Received by the editors July 1, 1980, and in final form August 15, 1980.
† Academia Sinica, Taipei, Taiwan.
‡ Bell Laboratories, Murray Hill, New Jersey 07974.
§ University of Massachusetts, Amherst, Massachusetts 01003.

nodes and links which connect the root to v, excluding v itself. The *length* of a path is the number of nodes on it. Node u is the *father* of node v, and v a *son* of u, if u has an outlink to v. Two nodes having the same father are called *brothers.*

A group testing algorithm can be represented by a binary tree where each internal node is associated with a test and its two outlinks are associated with the two possible outcomes. The *test history* at node v is the set of tests and outcomes associated with the nodes and links on the path for v.

Let D denote the set of the d defectives in the population. Any subset of the population is called a *sample point* of D if D can turn out to be s. For our problem, s is a sample point if and only if the cardinality of s is d. Associated with each node v is the set of sample points which are consistent with the test history of v. We refer to this set as the *sample space* at v and denote it by $S(v)$. Note that if v is a terminal point, then the cardinality of $S(v)$ is necessarily unity. We let $s(v)$ denote the sample point associated with the terminal node v.

When an algorithm T is in its binary tree representation, $M_T(n, d)$ is simply the maximum path length of the tree. Let $M_T(S)$ denote the maximum number of tests for the algorithm T to identify D from the sample space S, and define

$$M(S) = \min_T M_T(S).$$

The following two lemmas are straightforward and need no proofs.

LEMMA 1. *Suppose* $S_1 \subseteq S_2$. *Then* $M(S_1) \leqq M(S_2)$.

LEMMA 2. $M(S) \geqq \lceil \log_2 |S| \rceil$, *where* $|S|$ *is the cardinality of* S *and* $\lceil x \rceil$ *denotes the smallest integer not less than* x.

COROLLARY. $M(n, d) \geqq \lceil \log_2 \binom{n}{d} \rceil$.

3. The main results. Let $M(k|n, d)$ denote the minimax number of tests to identify the d defectives from n items when a particular set of k items is known to be contaminated.

LEMMA 3. $M(k|n, d) \geqq 1 + M(n-1, d-1)$ *for* $k \geqq 2$ *and* $n > d > 0$.

Proof. Without loss of generality, let $x_1, x_2, \cdots, x_k$ denote the k items in the contaminated set. Then for $k \geqq 2$,

$$M(k+1|n, d) \geqq M(k|n, d) \quad \text{by Lemma 1.}$$

Let T be any algorithm for the $(2|n, d)$ problem, and let m denote the maximum length of a path from the root of T to a terminal vertex; i.e., $m = M_T(2|n, d)$.

CLAIM. *Every path of length* m *in* T *includes at least one test that contains an item from the given contaminated set* $\{x_1, x_2\}$.

Proof of claim. Suppose, to the contrary, that for some terminal vertex v the path $p(v)$ has length m and includes no test containing x_1 or x_2. Since no test on $p(v)$ can distinguish between x_1 and x_2, and since $\{x_1, x_2\}$ is known to be contaminated, we conclude that x_1 and x_2 are both defective in the sample point $s(v)$. Let u be the brother node of v, which must also be a terminal node since v has maximum path length. The test history of u differs from that for v only in the outcome of the last test, so x_1 and x_2 must also both be defective in $s(u)$. Therefore, since $s(u) \neq s(v)$, there exist indices i and j such that x_i is defective and x_j is good in $s(u)$, while x_i is good and x_j is defective in $s(v)$ ($s(u)$ and $s(v)$ may also differ on other items.) Let f denote the father node of u and v, so that $S(f) = \{s(u), s(v)\}$. Then no test on the path $p(f)$ can have the form $G \cup \{x_j\}$ where G, possibly empty, contains only items classified as good in $s(u)$; such a test must have a good outcome for the sample point $s(u)$ and a contaminated outcome for the sample point $s(v)$, and hence would have separated $s(v)$ from $s(u)$. Define s to be the sample

point identical to $s(u)$ except that x_2 is good and both x_i and x_j are defective. The sample point s can be distinguished from $s(u)$ only by a test involving x_2, which by assumption does not exist on $p(f)$, or by a test of the form $G \cup \{x_j\}$ as described above, which we have also seen cannot occur on $p(f)$. Thus the sample point s must also belong to $S(f)$, and we cannot have both u and v being terminal nodes, a contradiction that completes the proof of the claim.

Without loss of generality, we may in fact assume that every path of length m in T includes at least one test that contains x_1. This follows by a simple relabeling; if the first test containing x_1 or x_2 on such a path does not contain x_1, we can interchange the names x_1 and x_2 in that test and all tests in the subtree below it without affecting the testing procedure, because this is merely renaming two items that have not yet been distinguished from one another.

We now apply this modified version of T to the $(n-1, d-1)$ problem, adding a known defective, labeled x_1, to the population and skipping all group tests involving x_1, since we already know the outcome of such a test. Because every path of maximum length m in T includes at least one such test, this procedure never uses more than $m-1$ tests, and we have

$$M_T(2|n, d) \geqq 1 + M_T(n-1, d-1),$$

from which Lemma 3 follows immediately.

We now give a new proof of a result reported in [1].

THEOREM 1. *$M(n, d) = n-1$ for $2d+1 \geqq n > d$.*

Proof. We prove Theorem 1 by induction on $n+d$. Theorem 1 is trivially true for $n+d=1$. To prove the general case, let T denote a minimax algorithm for the (n, d) problem. Suppose T first tests a set of k items. If $k>1$, then

$$\begin{aligned} M_T(n, d) &= 1 + \max\{M(n-k, d), M(k|n, d)\} \\ &\geqq 1 + M(k|n, d) \\ &\geqq 2 + M(n-1, d-1) \quad \text{by Lemma 3} \\ &= 2 + n - 2 = n \quad \text{by induction,} \end{aligned}$$

except when $n = 2d+1$; in that case

$$\begin{aligned} M_T(n, d) &\geqq 2 + M(2d, d-1) \\ &\geqq 2 + M(2d-1, d-1) \quad \text{by Lemma 1} \\ &= 2 + n - 3 = n-1 \quad \text{by induction.} \end{aligned}$$

If $k=1$, then

$$\begin{aligned} M_T(n, d) &= 1 + \max\{M(n-1, d), M(n-1, d-1)\} \\ &= 1 + (n-2) = n-1 \quad \text{by induction,} \end{aligned}$$

since at least one of the two sets of parameters $(n-1, d)$ and $(n-1, d-1)$ satisfies the conditions of Theorem 1. The proof is complete.

Next we prove a monotonicity property of $M(n, d)$.

THEOREM 2. *$M(n, d) \geqq 1 + M(n-1, d-1) \geqq M(n, d-1)$ for $n > d > 0$.*

Proof. The first inequality follows immediately from Lemma 3 if we note $M(n, d) = M(n|n, d)$. The second inequality is trivially true for $d=1$. We prove the general case by using the induction assumption

$$M(n-1, d-1) \geqq M(n-1, d-2).$$

Let T be an algorithm which first tests a single item and then uses a minimax algorithm for the remaining problem. Then

$$\begin{aligned} M(n, d-1) &\leqq M_T(n, d-1) \\ &= 1 + \max\{M(n-1, d-1), M(n-1, d-2)\} \\ &= 1 + M(n-1, d-1) \quad \text{by induction.} \end{aligned}$$

The proof is complete.

COROLLARY. *Suppose* $n-d>1$. *Then* $M(n, d)=n-1$ *implies* $M(n-1, d)=n-2$.

Proof. Suppose to the contrary that $M(n-1, d)<n-2$. Let T denote an algorithm for the (n, d) problem which first tests a single item and then uses a minimax algorithm for the remaining problem. Then

$$\begin{aligned} M(n, d) \leqq M_T(n, d) &= 1 + \max\{M(n-1, d), M(n-1, d-1)\} \\ &= 1 + M(n-1, d) \quad \text{by Theorem 2} \\ &< n-1, \end{aligned}$$

a contradiction to the assumptions of the corollary.

THEOREM 3. $M(n, d)=M(n, d-1)$ *implies* $M(n, d)=n-1$.

Proof. Suppose $n-d=1$. Then Theorem 3 follows from Theorem 1. We prove the general case by induction on $n-d$. Note that $M(n, d)=M(n, d-1)$ implies

$$M(n, d) = 1 + M(n-1, d-1) \quad \text{by Theorem 2.}$$

Let T be a minimax algorithm for the (n, d) problem. Suppose T first tests a set of k items. If $k>1$, then

$$\begin{aligned} M_T(n, d) &\geqq 1 + M(k|n, d) \\ &\geqq 2 + M(n-1, d-1) \quad \text{by Lemma 3,} \end{aligned}$$

a contradiction to what we just observed. Therefore $k=1$ and

$$\begin{aligned} M(n, d) &= 1 + \max\{M(n-1, d), M(n-1, d-1)\} \\ &= 1 + M(n-1, d), \end{aligned}$$

by Theorem 2 and the fact $d<n-1$. It follows that

$$\begin{aligned} &M(n-1, d-1) = M(n-1, d), \quad \text{hence} \\ &M(n-1, d-1) = n-2 \quad \text{by induction.} \end{aligned}$$

Therefore

$$M(n, d) = 1 + M(n-1, d-1) = n-1.$$

LEMMA 4. *Suppose* $M(n, d)<n-1$. *Then* $M(n, d) \geqq 2l + M(n-l, d-l)$ *for* $n>d \geqq l>0$.

Proof. We may assume $n-d>1$, for otherwise $M(n, d)=n-1$ by Theorem 1. We first prove Lemma 4 for $l=1$.

Let T be a minimax algorithm for the (n, d) problem which first tests a set of k items. If $k > 1$, then Lemma 4 is an immediate consequence of Lemma 3, as we showed in the proof of Theorem 3. Therefore we assume $k = 1$. Suppose to the contrary that

$$M_T(n, d) < 2 + M(n-1, d-1).$$

Then

$$\begin{aligned} 1 + M(n-1, d-1) &\geqq M_T(n, d) \\ &= 1 + \max\{M(n-1, d), M(n-1, d-1)\} \\ &= 1 + M(n-1, d) \quad \text{by Theorem 2.} \end{aligned}$$

Therefore

$$M(n-1, d-1) = M(n-1, d) = n-2 \quad \text{by Theorem 3.}$$

Consequently,

$$M(n, d) = 1 + M(n-1, d) = n-1,$$

a contradiction to the assumptions of Lemma 4. Thus Lemma 4 holds for the case $l = 1$. The general case is then proved by a straightforward induction argument (on l).

COROLLARY. $M(n, d) \geqq \min\left\{n-1, 2l + \left\lceil \log_2 \binom{n-l}{d-l} \right\rceil\right\}$ *for* $n > d \geqq l > 0$.

Define

$$f(k) = \binom{4k+1}{k} \Big/ \binom{4(k-1)+1}{k-1} \quad \text{for } k = 1, 2, \cdots.$$

LEMMA 5. $f(k) > f(k-1)$ *for* $k \geqq 2$.

Proof.

$$\begin{aligned} \frac{f(k)}{f(k-1)} &= \frac{\binom{4k+1}{k}\binom{4(k-2)+1}{k-2}}{\binom{4(k-1)+1}{k-1}^2} \\ &= \frac{(4k+1)(3k-2)}{(3k+1)(4k-3)} \cdot \frac{(4k-1)(3k-3)}{3k(4k-5)} \cdot \frac{(4k-2)(3k-4)}{(3k-1)(4k-6)} > 1, \end{aligned}$$

since each of the three factors is greater than one.

COROLLARY. $\binom{4k+1}{k} > 2^{3k-1}$ *for* $k \geqq 0$.

Proof. The corollary can be easily verified for $k \leqq 3$. Furthermore $f(4) = \binom{17}{4}/\binom{13}{3} = \frac{2380}{286} > 2^3$. Therefore $f(k) > 2^3$ for $k \geqq 4$ by Lemma 5. Using induction on k, we obtain the corollary.

THEOREM 4. $M\left(\left\lceil \frac{5d+1}{2} \right\rceil, d\right) = \left\lceil \frac{5d+1}{2} \right\rceil - 1.$

Proof. We obtain

$$M\left(\left\lceil\frac{5d+1}{2}\right\rceil, d\right) \geqq \min\left\{\left\lceil\frac{5d+1}{2}\right\rceil - 1, 2\left\lceil\frac{d}{2}\right\rceil + \left\lceil \log_2 \begin{pmatrix} \left\lceil\frac{5d+1}{2}\right\rceil - \left\lceil\frac{d}{2}\right\rceil \\ d - \left\lceil\frac{d}{2}\right\rceil \end{pmatrix} \right\rceil\right\}$$

$$\geqq \min\left\{\left\lceil\frac{5d+1}{2}\right\rceil - 1, 2\left\lceil\frac{d}{2}\right\rceil + \left\lceil \log_2 \begin{pmatrix} 4\left\lceil\frac{d-1}{2}\right\rceil + 1 \\ \left\lceil\frac{d-1}{2}\right\rceil \end{pmatrix} \right\rceil\right\}$$

$$\geqq \min\left\{\left\lceil\frac{5d+1}{2}\right\rceil - 1, 2\left\lceil\frac{d}{2}\right\rceil + 3\left\lceil\frac{d-1}{2}\right\rceil\right\}$$

$$= \left\lceil\frac{5d+1}{2}\right\rceil - 1,$$

by using first the corollary of Lemma 4 with $l = \lceil d/2 \rceil$, and then the corollary of Lemma 5.

COROLLARY. *$M(n, d) = n - 1$ for $\left\lceil\frac{5d+1}{2}\right\rceil \geqq n > d > 0$.*

Proof. It follows from Theorem 4 and the corollary of Theorem 2.

4. A conjecture. We make the following conjecture.

CONJECTURE. *$M(n, d) = n - 1$ for $3d \geqq n > d > 0$.*

If the conjecture is true, then it will be the sharpest result of this type, since we have

THEOREM 5. *$M(n, d) < n - 1$ for $n > 3d \geqq 3$.*

Proof. Theorem 5 can be easily verified for $d = 1$. We prove the general case by induction on d.

Let T be an algorithm for the (n, d) problem which can be described by the following steps.

Step 1. Set $i = 1$. Set $j = k = 0$.

Step 2. If $i > n$, stop. If $i = n$, test item x_n and stop.

Step 3. If $i < n$, test two items x_i and x_{i+1}. If the outcome is pure, set $i = i + 2$, $j = j + 1$. If the outcome is defective, set $k = k + 1$ and test x_i. If x_i is good (then x_{i+1} is defective), set $i = i + 2$; if x_i is defective, set $i = i + 1$.

Step 4. If $j \geqq k$, use a minimax algorithm for the remaining problem. If $j < k$, go back to Step 2.

Note that whenever Step 4 is executed, then at most $j + 2k$ tests have been taken with at least $2j$ good items and k defectives identified. If $j = k$, then

$$M_T(n, d) \leqq 3k + M(n - 3k, d - k)$$
$$< 3k + n - 3k - 1 = n - 1 \quad \text{by induction.}$$

If $j > k$, then

$$M_T(n, d) \leqq j + 2k + M(n - 2j - k, d - k)$$
$$\leqq j + 2k + (n - 2j - k - 1)$$
$$< n - 1.$$

If Step 4 is never executed, then all defectives are identified in at most $j+2d$ tests with $j<k=d$. Therefore

$$M_T(n, d) \leqq j+2d<3d \leqq n-1.$$

The proof is complete.

Acknowledgment. The authors wish to thank M. R. Garey for a careful reading and some useful suggestions.

REFERENCE

[1] F. K. HWANG, *A minimax procedure on group testing problems*, Tamkang J. Math., 2 (1971), pp. 39–44.

SIAM. J. ALG. DISC. METH.
Vol. 2, No. 2, June 1981

0196-5212/81/0202-0002 $01.00/0

AN $O(n^2)$ ALGORITHM FOR COLORING PROPER CIRCULAR ARC GRAPHS*

JAMES B. ORLIN†, MAURIZIO A. BONUCCELLI‡ AND DANIEL P. BOVET§

Abstract. A graph is a circular arc graph if each vertex of the graph is associated with an arc on a circle in such a way that two vertices of the graph are adjacent if and only if the corresponding arcs overlap. A circular arc graph is proper if none of the representing arcs is contained within another. An $O(n^2)$ algorithm is given for determining whether a proper circular arc graph with n nodes may be colored with k colors.

1. Introduction. A *circular arc family* is a set $F = \{A_1, \cdots, A_n\}$ of arcs on a circle. A circular arc family is *proper* if no arc is contained within another. A graph is a (*proper*) *circular arc graph* if there is a 1:1 correspondence between the vertices of the graph and the arcs of a (proper) circular arc family such that two vertices of the graph are adjacent if and only if the corresponding arcs overlap. For example, the graph in Fig. 1a is a proper circular arc graph, and Fig. 1b gives a proper circular arc model of this graph. The diagram in Fig. 1c is also a circular arc model; however, it is not proper because arc A_2 is contained in arc A_1.

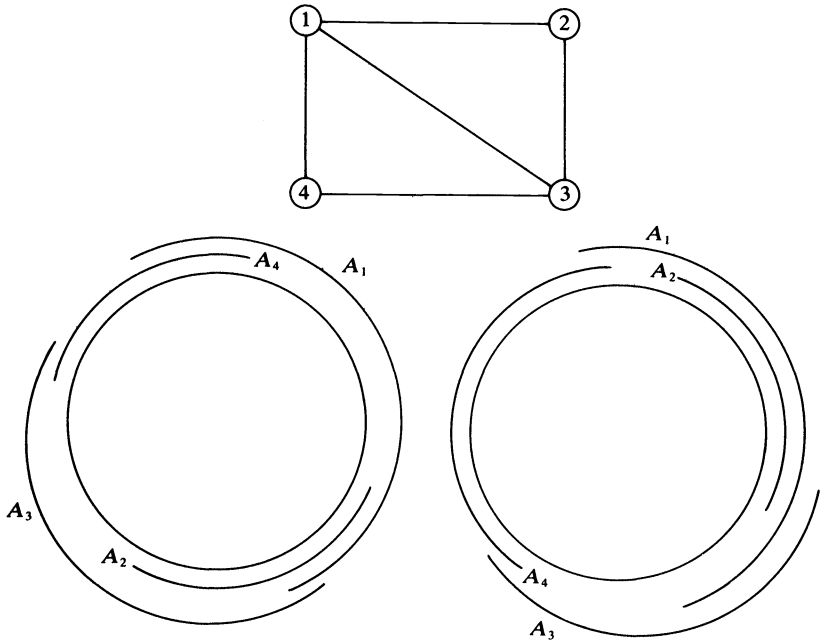


FIG. 1. a) *A proper circular arc graph.* b) *A proper circular arc representation for the graph in* a). c) *A (nonproper) circular arc representation for the graph in* a).

Circular arc graphs have been studied extensively. Tucker [12] has recently given a polynomial algorithm for recognizing these graphs. Gavril [6], [7] has given polynomial algorithms for finding a maximum independent set, a maximum clique and a minimum covering by cliques for circular arc graphs. The problem of coloring circular arc graphs

* Received by the editors April 1, 1980, and in revised form June 30, 1980.

† Alfred P. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

‡ Istituto di Scienze del'Informazione, Università di Pisa, Curso Italia 40, 56100 Pisa, Italy.

§ Istituto Matematico Guido Castenuovo, Università di Roma, Città Universitaria, 01000 Rome, Italy.

has been investigated by Tucker [11], and this problem was recently proved to be NP-complete by Garey, Johnson, Miller and Papadimitriou [5]. In this latter paper the problem of coloring proper circular arc graphs was mentioned as being a significant open problem. In [3] the problem of selecting the minimum number of node disjoint paths in a circular arc graph has been solved with an $O(n \log n)$ algorithm.

Applications. Coloring circular arc graphs has applications in both cyclic scheduling and in optimal register allocation in computer programs. Both of these applications are discussed in [11]. In cyclic scheduling we consider a number of tasks that have to be carried out periodically, and each arc represents a span of time during which the task is executed. For example, consider a limousine service at an airport that repeats its schedule every hour. Each arc represents the portion of the hour devoted to a specific (hourly repeated) round trip. The circular arc graph may be k-colored if and only if the corresponding limousine schedule can be serviced by k limousines such that each route is serviced periodically by the same limousine (see Fig. 2).

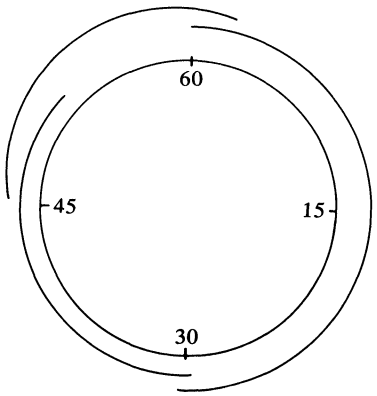


FIG. 2. *Arcs representing the timetables for three hourly repeated round trips leaving on the hour, half-past the hour, and three-quarters past the hour.*

For the computer application, consider a loop in a computer program and regard the flow of control in the loop as a circle. Each variable within the loop has a certain lifetime which may be modeled as an arc of the circle. Since it is necessary to store a variable only during its lifetime, a single register may store several variables as long as the lifetimes of any two of these variables do not intersect. There is a k-coloring of the corresponding circular arc graphs if and only if it is possible to store all the variables in k registers so that each variable is in only one register during its lifetime.

In the first application there is a restriction that each route must be traveled periodically by the same limousine. In the second application there is a restriction that each variable is stored repeatedly in the same register. If these restrictions are relaxed, the problem may be modeled as a coloring problem on "periodic interval graphs," which in turn is a special case of the periodic Dilworth's theorem, as formulated and solved in [8]. Furthermore, the problem is efficiently solvable even when one interval may contain another. Tucker [9] characterized proper circular arc graphs. A subclass which arises commonly in applications is that subclass induced by a family of arcs each with a common length.

2. Determining k-colorings for proper circular arc graphs.

Circular arc representations. Let G be a proper circular arc graph. Tucker [9] gives an efficient algorithm for creating a circular arc representation for these graphs, which runs in $O(n^2)$ time using as a subroutine the Booth and Lueker algorithm [4] for recognizing the consecutive ones property in matrices. In the following we will assume that a given proper circular arc graph G has an associated proper circular arc representation $F=\{A_1, \cdots, A_n\}$ such that $A_i=[a_i, b_i]$ and $a_i, b_i \in [0, 1)$. The interpretation is that A_i is the arc on the unit circle that stretches clockwise from point a_i to point b_i and contains both of these points. We also denote the graph as $G(F)$.

We also assume that the arcs are ordered so that $a_1 < a_2 < \cdots < a_n$, and we assume that there are at least two arcs which do not overlap (otherwise, the coloring is trivial). In the following the vertex set of G is $V=\{1, \cdots, n\}$.

Overlap cliques. An *overlap clique* of G is a maximal set of vertices of G whose corresponding arcs all intersect at a common point of the circle. It is easy to see that each overlap clique is induced by one of the points in the set $\{a_1, \cdots, a_n, b_1, \cdots, b_n\}$. Thus there are at most $2n$ such cliques.

Set S is called *circularly consecutive* if either $S=\{i, i+1, \cdots, j\}$ or else $S=\{i, i+1, \cdots, n, 1, \cdots, j\}$ for some $i, j \in \{1, \cdots, n\}$.

LEMMA 1. *If $G=G(F)$ is a proper circular arc graph, then each overlap clique is circularly consecutive.*

Proof. If the point inducing the overlap clique is $p=0$, then $S=\{i: a_i > b_i \text{ or } a_i = 0\}$, and this set is circularly consecutive. Let $x \pmod 1$ denote the fractional part of x. For $p \neq 0$, the overlap clique induced by p is $s=\{i: (a_i-p) \pmod 1 = (b_i-p) \pmod 1 \text{ or } a_i \pmod 1 = p\}$, which is circularly consecutive. □

For a given circularly consecutive set S, the *last element* of S is the unique element $i \in S$ such that $i+1 \notin S$ (the last element of S is n if $n \in S$ and $1 \notin S$). Henceforth, we will write an overlap clique as $S=\langle i, j\rangle$ where j is the last element of S and i is the last element of $\{i, \cdots, n\}-S$. For example, if $n=5$, then $\{3, 4, 5\}=\langle 2, 5\rangle$, and $\{4, 5, 1\}=\langle 3, 1\rangle$. This choice of our representation will be made more clear in the context of the algorithm.

LEMMA 2. *Let $G=G(F)$ be a circular arc graph with n vertices, and let k be a divisor of n. Then G is k-colorable if and only if G has no overlap clique of $k+1$ vertices.*

Proof. The "only if" part is trivial, since no graph with a clique of $k+1$ vertices is k-colorable. For the "if" direction, consider the coloring of G with colors $0, 1, \cdots, k-1$ such that vertex i is assigned color $i \pmod k$. If $i<j$ and nodes i and j are assigned the same color, then either $a_i \in [a_i, b_i]$ or else $b_j \in [a_i, b_i]$. In the former case vertices $i, i+1, \cdots, j$ are in the same overlap clique; in the latter case vertices $j, \cdots, n, 1, \cdots, i$ are in the same overlap clique. In both cases the overlap clique has at least $k+1$ vertices. □

In the following, $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ denote the least integer function and the greatest integer function.

LEMMA 3. *Let G be a proper circular arc graph that is k-colorable. Then G may be k-colored in such a way that each color class has either $\lceil n/k \rceil$ or $\lfloor n/k \rfloor$ colors.*

Proof. We first show the result for $k=2$. If n is even and $k=2$, the result is a consequence of the coloring given in Lemma 2. Suppose there is a 2-coloring of G, and assume n is odd. One of the color classes has at least $(n+2)/2$ vertices; otherwise, there is nothing to prove. This color class contains vertices i and $i+1$ for some i. Consider now the subset

$$C=\{i, i-2, i-4, \cdots\} \cup \{i+1, i+3, i+5, \cdots\}.$$

Neither C nor $V-C$ has two adjacent vertices; else it would imply a clique in G of size 3. Thus there is a 2-coloring with color classes C and $V-C$ with $(n+1)/2$ and $(n-1)/2$ vertices, thus proving the lemma for the case that $k=2$.

Consider now a k-coloring for $k>2$, and let C and D be color classes with the greatest and least number of vertices. By the above, we may partition the set $C \cup D$ into two color classes whose cardinality differs by at most one. Iteratively applying this recoloring procedure we obtain a coloring in which each color class has $\lceil n/k \rceil$ or $\lfloor n/k \rfloor$ vertices, proving the lemma. □

THEOREM 1. *Let $G = G(F)$ be a proper circular arc graph with n vertices. Let k be an integer less than n, and let $r = n \pmod k$ with $0 \leqq r \leqq k-1$. Graph G may be k-colored if and only if there exists a subset V' of $r \cdot \lceil n/k \rceil$ vertices such that* (1) *the subgraph of G induced by V' has no overlap clique of size $r+1$, and* (2) *the subgraph of G induced by $\{1, \cdots, n\} - V'$ has no overlap clique of size $k-r+1$.*

Proof. If k is a divisor of n, the theorem follows from Lemma 2. Suppose there is a k-coloring of G, and suppose that $r = n \pmod k \neq 0$. By Lemma 3, there is a k-coloring such that each color class has $\lceil n/k \rceil$ or $\lfloor n/k \rfloor$ vertices. There are r color classes of size $\lceil n/k \rceil$ since the number of vertices of G is n. Let V' be the union of these color classes. The subgraph induced by V' is r-colorable and thus has no overlap clique of size $r+1$. The subgraph induced by $\{1, \cdots, n\} - V'$ is $(k-r)$-colorable and hence has no overlap clique of size $k-r+1$.

Conversely, suppose there is a subset V' of vertices satisfying the conditions of this theorem. The subgraph of G induced by V' has $r \cdot \lceil n/k \rceil$ vertices and may be r-colored by Lemma 2. The subgraph of G induced by $\{1, \cdots, n\} - V'$ has $(k-r) \cdot \lfloor n/k \rfloor$ vertices and may be $(k-r)$-colored by Lemma 2. Hence G may be k-colored. □

There is no a priori reason why the above partition problem should appear easier to solve than the coloring problem; however, the partition problem is really a special case of the shortest-path problem, a consequence of the following integer programming formulation of the partition problem. A similar partition problem for proper circular arc graphs was solved in a similar way by Bartholdi, Orlin and Ratliff in [1].

In the following we wish to determine a subset V' of vertices of G satisfying the conditions of Theorem 1. We let $x_i = |\{1, \cdots, i\} \cap V'|$, which is the number of vertices with index at most i in the set V'. Thus $i \in V'$ if and only if $x_i - x_{i-1} = 1$, where $x_0 = 0$. As before, we let $r = n \pmod k$. We now wish to determine a feasible solution to the system of constraints (1):

$$0 \leqq x_i - x_{i-1} \leqq 1 \quad \text{for } i = 1, \cdots, n, \tag{1a}$$

$$x_0 = 0, \tag{1b}$$

$$x_n = r \cdot \lceil n/k \rceil. \tag{1c}$$

For each overlap clique $S = \langle i, j \rangle$ with $i < j$,

$$x_j - x_i \leqq r, \tag{1d}$$

$$|S| - (x_j - x_i) \leqq k - r. \tag{1e}$$

Finally, for each overlap clique $S = \langle i, j \rangle$ with $i > j$,

$$x_j - x_i + x_n \leqq r, \tag{1f}$$

$$|S| - (x_j - x_i + x_n) \leqq k - r \tag{1g}$$

and

(1h) x_i is integer valued for $i = 1, \cdots, n$.

THEOREM 2. *Circular arc graph G may be k-colored if and only if there is a feasible solution to system* (1). *Such a solution may be determined in $O(n^2)$ steps.*

Proof. We note first that there is a 1:1 correspondence between subsets V' of vertices of G and integer vectors $x = (x_i)$ satisfying (1a) and (1b). The correspondence is given by the relation $x_i = |V' \cap \{1, \cdots, i\}|$. With this correspondence, each overlap clique $S = \langle i, j \rangle$ with $i < j$ is such that $|S \cap V'| = x_j - x_i$. If $S = \langle i, j \rangle$ with $i > j$, then $|S \cap V'| = x_j - x_i + x_n$.

We interpret the constraints of (1) as follows: (1c) requires that V' has $r \cdot \lceil n/k \rceil$ vertices; (1d) and (1f) require that each overlap clique in the subgraph induced by V' has at most r vertices; (1e) and (1g) require that each overlap clique in the subgraph induced by $\{1, \cdots, n\} - V'$ has at most $k - r$ vertices. Thus (1) is equivalent to requiring that G satisfy the conditions of Theorem 1.

Since x_n is fixed in value in the constraint (1c), we may eliminate x_n from the constraints (1f) and (1g). Each of the resulting constraints may be written as $x_j - x_i \leqq d_{ij}$ for an appropriate value d_{ij} (where i or j may be 0). Consider now a directed graph G' with vertex set $\{0, \cdots, n-1\}$, and for each constraint "$x_j - x_i \leqq d_{ij}$" of (1), there is an associated edge (i, j) of G' with distance d_{ij}. Then a feasible solution for (1) is $x'_0, \cdots, x'_{n-1}, x'_n$, where x'_j is the minimum distance in G' from node 0 to node j for $j = 0, \cdots, n-1$, and $x'_n = r \cdot \lceil n/k \rceil$. Let m denote the number of edges of G'. Then these distances may be computed in $O(n \cdot m)$ steps by the Bellman–Ford method [2], which in turn is $O(n^2)$ steps because each edge is associated with an inequality of (1), and there are at most $2n$ overlap cliques.

Once a feasible solution for (1) is determined, the coloring may be carried out in $O(n)$ steps via Lemma 2. □

COROLLARY 5. *A minimum coloring for a circular arc graph may be determined in $O(n^2 \log n)$ steps.*

Proof. It suffices to use binary search to determine the minimum value of k for which the given graph is k-colorable. This takes $O(\log n)$ iterations, each with $O(n^2)$ steps. □

The network G' of the proof of Theorem 1 is highly structured. An open question is whether the shortest-path distances may be computed faster than $O(n^2)$ using a specialized algorithm.

Acknowledgment. We gratefully acknowledge the helpful comments of Alan Tucker.

REFERENCES

[1] J. J. BARTHOLDI, III, J. B. ORLIN AND H. D. RATLIFF, *Cyclic scheduling via integer programs with circular ones*, Operations Research, 28 (1980), pp. 1074–1085.

[2] R. E. BELLMAN, *On a routing problem*, Quart. Appl. Math., 16 (1958), pp. 87–90.

[3] M. A. BONUCCELLI AND D. P. BOVET, *Minimum node disjoint path covering for circular-arc graphs*, Inform. Process Lett., 8 (1979), pp. 159–161.

[4] K. S. BOOTH AND G. S. LUEKER, *Linear algorithms to recognize interval graphs and test for consecutive ones property*, Proc. 7th Annual ACM Symposium on the Theory of Computing, New York, 1975, pp. 255–265.

[5] M. R. GAREY, D. S. JOHNSON, G. L. MILLER AND C. H. PAPADIMITRIOU, *The complexity of coloring circular arcs and chords*, this Journal, 1 (1980), pp. 216–227.

[6] F. GAVRIL, *Algorithms for a maximum clique and a maximum independent set of a circle graph*, Networks, 3 (1973), pp. 261–273.
[7] ———, *Algorithms on circular-arc graphs*, Networks, 4 (1974), pp. 357–369.
[8] J. B. ORLIN, *Periodic Dilworth's theorem with applications to cyclic scheduling*, work in progress.
[9] A. TUCKER, *Matrix characterization of circular-arc graphs*, Pacific J. Math., 39 (1971), pp. 535–545.
[10] ———, *Structure theorems for some circular-arc graphs*, Discrete Math., 7 (1974), pp. 167–195.
[11] ———, *Coloring a family of circular arcs*, SIAM J. Appl. Math., 29 (1975), pp. 493–502.
[12] ———, *An efficient test for circular-arc graphs*, SIAM J. Comput. 9 (1980), pp. 1–24.

SIAM. J. ALG. DISC. METH.
Vol. 2, No. 2, June 1981

0196-5212/81/0202-0003 $01.00/0

GENERALIZED SCHUR REPRESENTATION OF MATRIX-VALUED FUNCTIONS*

P. DELSARTE,† Y. GENIN†‡ AND Y. KAMP†

Abstract. The generalized Schur representation of a function matrix $\Omega(e^{i\theta})$ satisfying $\|\Omega\|_\infty \leqq 1$ is investigated in connection with certain results concerning the extensions of block-Hankel operators acting on Hilbert spaces. Various properties of such representations are elucidated, including a parametrization of $\Omega(e^{i\theta})$ in terms of a double sequence of Schur parameter matrices. Special attention is paid to the way in which the representation and parametrization of the shifted function $e^{ik\theta}\Omega(e^{i\theta})$ are related to those of $\Omega(e^{i\theta})$. In particular, the asymptotic behavior of the shifted representation for $k \to \pm\infty$ is studied in detail. The whole theory is developed so as to be of direct use in the analysis of half-plane block-Toeplitz systems.

1. Introduction. In a masterful paper, Adamjan, Arov and Krein [1] have brought to a high degree of achievement the theory of infinite Hankel matrices, and especially its relationship with various extension and approximation problems, including generalizations of the Schur [18], Takagi [19] and Nevanlinna–Pick [14], [16] problems.

Surprisingly enough, it turns out that parts of the same material play a central role in quite a different area of significant engineering interest. In two-dimensional digital filtering as well as stochastic estimation, techniques based upon recursive half-plane filtering and half-plane spectral factorization have recently received considerable attention (see, e.g., [4], [10], [13], [15]). To a great extent, these techniques fit into the framework of the theory of the so-called half-plane Toeplitz systems [4], [7], [8], [13]. As a matter of fact, although it has grown up quite independently, this theory has intimate connections with the problem of the extension of Hankel operators [1].

Some generalizations of their previous results to matrix-valued functions have been worked out by Adamjan, Arov and Krein [2]. In particular, a complete solution has been given to the problem of the extension of block-Hankel operators, leading to a well-defined representation in terms of two matrix-valued Schur functions.

It is natural to anticipate that, in a theory of half-plane block-Toeplitz systems, this representation will play the same illuminating role as in the scalar case. The precise aim of the present paper is to investigate in detail those properties of the above mentioned generalized Schur representation which are of direct applicability to the subject of half-plane block-Toeplitz systems [9]. This is the reason why special attention is paid to the so-called shift operation acting on the representation and to the related convergence properties, which occur as key issues in the analysis of half-plane Toeplitz systems [8].

In § 2, basic facts regarding matrix-valued Schur functions are first recalled [6]. Such a function $\Phi(z)$ can be parametrized by a well-defined sequence of matrices resulting from a matrix version of the classical Schur decomposition algorithm. It is then shown how a shifted copy $e^{-in\theta}\Phi(e^{i\theta})$ of the function $\Phi(e^{i\theta})$ can be represented by means of two Schur functions the parameter matrices of which are directly obtainable from those of Φ. The second part of § 2 is devoted to briefly describing the following

* Received by the editors February 26, 1980, and in revised form September 26, 1980.

† Philips Research Laboratory, Av. Van Becelaere 2, Box 8, B-1170, Brussels, Belgium.

‡ The work of this author was supported in part by the U.S. Army Research Office under contract DAAG-29-79-C-0215 and by the U.S. Air Force Office of Scientific Research A.F. Systems Command under contract AF-620-79-C-0058, while the author was on leave at Information Systems Laboratory, Stanford University.

result, due to Adamjan, Arov and Krein [2], which implies that the Schur representation exists for a much wider class of functions. Let there be given a block-Hankel matrix $\Gamma = [\Omega_{-s-t}: s \geqq 0, t \geqq 0]$ with norm $\|\Gamma\| < 1$ when acting on the Hilbert space l_2. Then the general matrix-valued function $\Omega(e^{i\theta})$ with Fourier coefficients Ω_{-t} for $t \geqq 0$ and subject to $\|\Omega\|_\infty = \operatorname{ess\,sup} \|\Omega(e^{i\theta})\| \leqq 1$ is represented by means of two Schur functions $\Phi(z)$ and $\Psi(z)$, where $\Psi(z)$ is determined from Γ while $\Phi(z)$ is arbitrary; more precisely, the functions Φ are in one-to-one correspondence with the extensions Ω of Γ.

The main theme of the present paper is introduced in § 3. It is concerned with the properties of the generalized Schur representation of a given function matrix $\Omega(e^{i\theta})$, of dimension $p \times p$, with $\|\Omega\|_\infty \leqq 1$, for which the associated block-Hankel operator Γ satisfies $\|\Gamma\| < 1$. A close relationship is first established between the desired representation of Ω, characterized by the Schur pair (Φ, Ψ), and a certain canonical factorization of the $2p \times 2p$ Hermitian matrix Δ with blocks $\Delta_{11} = \Delta_{22} = I$, $\Delta_{12} = \tilde{\Delta}_{21} = -\Omega$. As a result, a simple proof of the existence and uniqueness of the representation is obtained. The canonical factorization of Δ appears to be a key point in the argument showing how the Schur pair (Φ_{-k}, Ψ_k) of the k-shift $\Omega_k(e^{i\theta}) = e^{ik\theta}\Omega(e^{i\theta})$ is easily computable from (Φ, Ψ). In particular, for $k \geqq 1$, it is shown that the sequence of Schur parameters of Ψ_k is obtained from deleting the first k parameters $E_0, E_{-1}, \cdots, E_{1-k}$ of Ψ, while the sequence of Φ_{-k} results from adding $\tilde{E}_{1-k}, \cdots, \tilde{E}_{-1}, \tilde{E}_0$ in front of the sequence of Φ. An interpretation of this result is given in the framework of the theory of Szegö orthogonal polynomial matrices, which turns out to be useful in discussing convergence problems. The last topic treated in § 3 is the natural duality exchanging the roles of Φ and Ψ in the representation of a function Ω satisfying $\|\Omega\|_\infty < 1$. The dual function Ω', with Schur pair (Ψ, Φ), is explicitly identified, together with the corresponding canonical factorization.

Section 4 is devoted to certain convergence properties of the representation of the k-shifts of Ω when k goes to plus or minus infinity. Two types of convergence are examined. First, convergence in the mean is established, under the weak assumption $\|\Omega\|_\infty < 1$. Next, convergence in the sense of certain Wiener type algebras is mentioned, in connection with summability properties of the sequences of Schur parameters.

2. Preliminaries. This section reviews certain basic results concerning the parametrization of matrix-valued Schur functions [6], on the one hand, and the extension of infinite block-Hankel matrices [2], on the other hand. These results form the general background of our study.

2.1. Schur recurrence relations. Let $\Phi(z)$ be a $p \times p$ matrix all entries of which are analytic functions in the open unit disk $|z| < 1$. Then $\Phi(z)$ is said to belong to the class S (referring to Schur [18]) if the spectral norm $\|\Phi(z)\|$ does not exceed unity in $|z| < 1$. With the notation $A \leqq B$ meaning that $B - A$ is nonnegative definite, the required condition $\|\Phi(z)\| \leqq 1$ is equivalent to $\Phi(z)\tilde{\Phi}(z) \leqq I$ as well as to $\tilde{\Phi}(z)\Phi(z) \leqq I$.

The sequence of *Schur parameter matrices* $(F_1, F_2, \cdots)$ of a given function $\Phi_1(z) \in S$ are determined by $F_s = \Phi_s(0)$ together with the recurrence relation

$$\Phi_{s+1}(z) = z^{-1}(I - F_s\tilde{F}_s)^{-1/2}[\Phi_s(z) - F_s][I - \tilde{F}_s\Phi_s(z)]^{-1}(I - \tilde{F}_sF_s)^{1/2}, \tag{1}$$

for $s = 1, 2, \cdots$, with $X^{1/2}$ standing for the Hermitian square root of X. By definition, F_1 is *contractive*; i.e., $\|F_1\| \leqq 1$. In case of equality, $\|F_1\| = 1$, the function matrix $\Phi_1(z)$ is degenerate in the sense that it shrinks to a dimension $p' < p$. (Details can be found in [6].) If F_1 is strictly contractive, i.e., $\|F_1\| < 1$, then (1) yields a function matrix $\Phi_2(z) \in S$, and thus $\|F_2\| \leqq 1$. Iterating this process we obtain a sequence of strictly contractive matrices F_s, which may be either finite (degenerate case) or infinite (nondegenerate

case), together with a sequence of class S function matrices $\Phi_s(z)$. Conversely, at least in the nondegenerate case, $\Phi_1(z)$ can be uniquely reconstructed from the sequence of parameter matrices F_s. (There exists a similar but more complicated parametrization in the degenerate case [6].)

For future use let us define the symmetric permutation matrix W and the diagonal matrices J and $T(z)$, all of order $2p$, as follows:

$$W=\begin{bmatrix}0 & I\\ I & 0\end{bmatrix},\qquad J=\begin{bmatrix}I & 0\\ 0 & -I\end{bmatrix},\qquad T(z)=\begin{bmatrix}zI & 0\\ 0 & I\end{bmatrix}, \tag{2}$$

with 0 and I the $p\times p$ zero matrix and unit matrix, respectively. Next, given a strictly contractive $p\times p$ matrix E, define the $2p\times 2p$ Hermitian matrix

$$H(E)=\begin{bmatrix}(I-E\tilde{E})^{-1/2} & E(I-\tilde{E}E)^{-1/2}\\ \tilde{E}(I-E\tilde{E})^{-1/2} & (I-\tilde{E}E)^{-1/2}\end{bmatrix}. \tag{3}$$

This is the J-unitary version of the Halmos extension of $\tilde{E}$. In fact, $H(E)$ satisfies $\tilde{H}JH=J$ and thus is a *J-unitary matrix*. Other useful properties of (3) are

$$H(-E)=H(E)^{-1},\qquad H(\tilde{E})=WH(E)W. \tag{4}$$

In the sequel we shall often use the concept of *homographic transformation* (see especially Potapov [17]). Given a $2p\times 2p$ matrix M with $p\times p$ blocks $M_{ij}(i,j=1,2)$, the homographic image of a $p\times p$ matrix X under the action of M is defined to be $M[X]=(M_{11}X+M_{12})(M_{21}X+M_{22})^{-1}$. It is then easily verified that the inverse version of (1) can be written as

$$z\Phi_s(z)=T(z)H(F_s)[z\Phi_{s+1}(z)], \tag{5}$$

with $T(z)$ and $H(F_s)$ as in (2) and (3).

Let $T_s(z)=\prod_{t=1}^{s}T(z)H(F_t)$. Iterating (5) yields $z\Phi_1(z)=T_s(z)[z\Phi_{s+1}(z)]$. By definition, $T_s(e^{i\theta})$ is J-unitary and has the form

$$T_s(e^{i\theta})=T(e^{is\theta})\begin{bmatrix}\tilde{A}_s(e^{i\theta}) & \tilde{B}_s(e^{i\theta})\\ C_s(e^{i\theta}) & D_s(e^{i\theta})\end{bmatrix}, \tag{6}$$

where A_s, B_s, C_s and D_s are polynomial matrices of formal degree $s-1$. Define $\Psi_s(z)=B_s(z)A_s(z)^{-1}$. From the fact that $T_s(z)$ is J-contractive inside the unit circle (i.e., $\tilde{T}_s(z)JT_s(z)\leqq J$ for $|z|\leqq 1$) it follows that the rational function matrices A_s^{-1}, D_s^{-1} and Ψ_s belong to the class S. In addition, J-unitarity of $T_s(e^{i\theta})$ yields $\Psi_s(z)=D_s(z)^{-1}C_s(z)$. As a result, the formula $z\Phi_1=T_s[z\Phi_{s+1}]$ with $z=e^{i\theta}$ can be written as

$$e^{i(1-s)\theta}\Phi_1=\tilde{A}_s(\tilde{\Psi}_s+e^{i\theta}\Phi_{s+1})(I+e^{i\theta}\Psi_s\Phi_{s+1})^{-1}D_s^{-1}, \tag{7}$$

where the argument $e^{i\theta}$ is omitted. It is important to note that, except for normalization, A_s and D_s are uniquely determined from Ψ_s. Indeed, J-unitarity of $T_s(e^{i\theta})$ yields both spectral factorization formulas,

$$(I-\tilde{\Psi}_s\Psi_s)^{-1}=A_s\tilde{A}_s,\qquad (I-\Psi_s\tilde{\Psi}_s)^{-1}=\tilde{D}_sD_s, \tag{8}$$

on the unit circle. Let us point out, without going into details, that the matrices A_s, B_s, C_s and D_s have a direct interpretation in the framework of the theory of the Szegö orthogonal polynomial matrices [5], [6] (cf., § 3.2 below). This leads to the formula $z\Psi_k(z)=T(z)H(\tilde{F}_k)[z\Psi_{k-1}(z)]$ for $k=s,s-1,\cdots,1$, showing that the sequence of Schur parameters associated with $\Psi_s(z)$ is $(\tilde{F}_s,\tilde{F}_{s-1},\cdots,\tilde{F}_1,0,0,\cdots)$.

2.2. Extension of block-Hankel operators. It turns out that a representation of type (7), (8) can be exhibited for a class of function matrices considerably larger than that consisting of the shifted versions $e^{-in\theta}\Phi$ of functions $\Phi \in S$, as considered in § 2.1. Before entering the subject, let us explain how it fits into the theory of block-Hankel operators developed by Adamjan, Arov and Krein [2]. Let

$$\Gamma = \begin{bmatrix} \Omega_0 & \Omega_{-1} & \Omega_{-2} & \cdot \\ \Omega_{-1} & \Omega_{-2} & \cdot & \cdot \\ \Omega_{-2} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix} \tag{9}$$

be the infinite block-Hankel matrix built on a given sequence of $p \times p$ complex matrices $(\Omega_0, \Omega_{-1}, \cdots)$. Assume that Γ acts as a bounded operator on the Hilbert space l_2, with norm $\|\Gamma\| < 1$. Then the set of $p \times p$ function matrices $\Omega(e^{i\theta})$, subject to

$$\|\Omega\|_\infty = \operatorname{ess\,sup} \|\Omega(e^{i\theta})\| \leqq 1 \tag{10}$$

and having Ω_j as Fourier coefficient of index j (for $j = 0, -1, -2, \cdots$), is in one-to-one correspondence with the set of $p \times p$ function matrices $\Phi(z) \in S$ via the formula

$$\Omega = \tilde{A}(\tilde{\Psi} + e^{i\theta}\Phi)(I + e^{i\theta}\Psi\Phi)^{-1}D^{-1}, \tag{11}$$

where $A(z)$, $D(z)$ and $\Psi(z)$ are determined from Γ as explained below (see [2]).

It is convenient to view Γ as an operator mapping the space L_2^+ into the space L_2^-, where L_2^+ (resp. L_2^-) consists of the $p \times p$ function matrices with square integrable entries having vanishing Fourier coefficients of negative (resp. positive) indices. Thus, according to (9), the formula $\Gamma X = Y$, with $X \in L_2^+$ and $Y \in L_2^-$, means $\sum_{i=0}^{\infty} \Omega_{-i-j}X_i = Y_{-j}$ for $j = 0, -1, -2, \cdots$. Let Γ^* denote the adjoint of Γ, mapping L_2^- into L_2^+. Under the assumption $\|\Gamma\| < 1$, the system of equations

$$\begin{aligned} C - \Gamma^*\tilde{A} = 0, \qquad & \tilde{B} - \Gamma D = 0, \\ \tilde{A} - \Gamma C = A(0)^{-1}, \qquad & D - \Gamma^*\tilde{B} = \tilde{D}(0)^{-1} \end{aligned} \tag{12}$$

has a solution $A, B, C, D \in L_2^+$ (with $A(0) = A_0$ and $D(0) = D_0$ nonsingular). Moreover, this solution is unique except for substitutions of the form $A \to \tilde{U}A$, $B \to \tilde{V}B$, $C \to CU$, $D \to DV$, where U and V are constant unitary matrices. (Note that (12) can be solved by an exponentially convergent iterative method; cf. [7].) Define the $2p \times 2p$ function matrix

$$L(e^{i\theta}) = \begin{bmatrix} \tilde{A}(e^{i\theta}) & \tilde{B}(e^{i\theta}) \\ C(e^{i\theta}) & D(e^{i\theta}) \end{bmatrix}. \tag{13}$$

The main properties of the solution of (12) are expressed by the fact that L is J-unitary and that both A^{-1} and D^{-1} belong to L_2^+. As a consequence, the function matrices $A(z)^{-1}$, $D(z)^{-1}$ and $\Psi(z) = B(z)A(z)^{-1}$ belong to the class S, with the property $\Psi(z) = D(z)^{-1}C(z)$. In addition, $A(z)$ and $D(z)$ are the left and right spectral factors of $(I - \tilde{\Psi}\Psi)^{-1}$ and of $(I - \Psi\tilde{\Psi})^{-1}$, respectively, in the sense that they satisfy

$$(I - \tilde{\Psi}\Psi)^{-1} = A\tilde{A}, \qquad (I - \Psi\tilde{\Psi})^{-1} = \tilde{D}D \tag{14}$$

on the unit circle $z = e^{i\theta}$. One of the main results of [2] says that, when substituted into (11), the triple (A, D, Ψ) yields the parametric representation of all functions Ω with $\|\Omega\|_\infty \leqq 1$ having the prescribed Fourier coefficients $\Omega_0, \Omega_{-1}, \cdots$ appearing in Γ. Note that (11) can be written in terms of a homographic transformation as $\Omega = L[e^{i\theta}\Phi]$. Note also the equivalent formula $\tilde{\Omega} = WLW[e^{-i\theta}\tilde{\Phi}]$, with W as in (2).

3. Generalized Schur representation. We are now in a position to specify the subject of the present paper. Let there be given a $p \times p$ function matrix $\Omega(e^{i\theta})$ satisfying (10). One is looking for an expression of the form (11), where $\Phi(z)$ and $\Psi(z)$ are class S function matrices while $A(z)$ and $D(z)$ are outer function matrices of Hardy class H_2, related to $\Psi(z)$ by the spectral factorization conditions (14). In this case, (11) is called a *generalized Schur representation* of Ω and the corresponding pair (Φ, Ψ) is called a *Schur pair* of Ω. Note that $\Psi(z)$ uniquely specifies $A(z)$ and $D(z)$ within right and left unitary factors, respectively. The intrinsic properties of $\Psi \in S$ warranting the existence of suitable functions A and D amount to integrability of $\operatorname{tr}[(I-\Psi\tilde{\Psi})^{-1}]$ and $\log\det(I-\Psi\tilde{\Psi})$ on the unit circle. Such properties are generally not required for the companion function $\Phi(z)$.

Defining the block-Hankel matrix Γ from Ω as in (9), one clearly has $\|\Gamma\| \leqq \|\Omega\|_\infty$, hence $\|\Gamma\| \leqq 1$, by (10). The results of [2] reviewed in § 2.2 establish the existence of a generalized Schur representation of Ω in case of strict inequality, $\|\Gamma\| < 1$. In addition, it follows quite immediately from these results that the representation is unique (see the end of § 3.1). As shown by (7) and (8), a simple illustration of the theory is provided by shifted class S function matrices $\Omega = e^{-in\theta}\Phi_1$; this particular case corresponds to block-Hankel matrices (9) with a finite number of nonzero rows and columns. For future use, observe that the action of the operator Γ and of its adjoint Γ^* is described by

$$\begin{aligned} \Gamma X &= (\Omega X)^- \quad \text{for } X \in L_2^+, \\ \Gamma^* Y &= (\tilde{\Omega} Y)^+ \quad \text{for } Y \in L_2^-, \end{aligned} \tag{15}$$

where $(F)^\pm$ denotes the projection into $L_2^\pm$ of a given $p \times p$ function matrix F belonging to L_2.

3.1. A related factorization problem. From the generalized Schur representation (11) of Ω, construct the matrix $L(e^{i\theta})$ as in (13), with $B(z) = \Psi(z)A(z)$ and $C(z) = D(z)\Psi(z)$. It appears from (14) that $L(e^{i\theta})$ is J-unitary. Next, construct four $p \times p$ function matrices

$$\begin{aligned} P(z) &= A(z)^{-1}[I + z\Phi(z)\Psi(z)]^{-1}, \qquad Q(z) = P(z)\Phi(z), \\ S(z) &= [I + z\Psi(z)\Phi(z)]^{-1}D(z)^{-1}, \qquad R(z) = \Phi(z)S(z). \end{aligned} \tag{16}$$

In view of the fact that $(I + z\Phi\Psi)^{-1}$ and $(I + z\Psi\Phi)^{-1}$ belong to the class C (referring to Carathéodory; see, e.g., [6]) it follows that P, Q, R and S are Hardy functions. Note that $A(0)P(0) = S(0)D(0) = I$. From (16) construct the $2p \times 2p$ matrix

$$K(e^{i\theta}) = \begin{bmatrix} P(e^{i\theta}) & -e^{i\theta}Q(e^{i\theta}) \\ -e^{-i\theta}\tilde{R}(e^{i\theta}) & \tilde{S}(e^{i\theta}) \end{bmatrix}. \tag{17}$$

Let us now derive a useful relationship between K and L, which plays an important role in our study. Define

$$\Delta(e^{i\theta}) = \begin{bmatrix} I & -\Omega(e^{i\theta}) \\ -\tilde{\Omega}(e^{i\theta}) & I \end{bmatrix}. \tag{18}$$

Writing (11) in terms of homographic transformations, as mentioned at the end of § 2.2, one immediately obtains from (16) the remarkable identity

$$\Delta = K(J\tilde{L}J) = KL^{-1}. \tag{19}$$

In particular, $K = \Delta L$ implies that P, Q, R and S necessarily belong to the Hardy class H_2 (isomorphic to L_2^+).

It turns out that the properties mentioned above entirely characterize the generalized Schur representation. In fact, it is shown in Theorem 1 below that, provided $\|\Gamma\| < 1$, a factorization (19) of $\Delta(e^{i\theta})$, with suitable matrices $K(e^{i\theta})$ and $L(e^{i\theta})$, exists and is unique within normalization. Moreover this factorization, which will be referred to in the sequel as the *canonical factorization* of Δ, directly produces the unique generalized Schur representation of Ω. Besides its own interest, the result of Theorem 1 is important to this paper because it leads to simple proofs of the shift and duality properties (see §§ 3.2 and 3.3). Let us stress that the existence and uniqueness of the generalized Schur representation belong to Adamjan, Arov and Krein [2]. However, the proof given here does not resort to the very general theory developed in [2].

THEOREM 1. *Let Ω be a function matrix satisfying $\|\Omega\|_\infty \leqq 1$ and $\|\Gamma\| < 1$. Then there exists a factorization $\Omega = KL^{-1}$, where the $p \times p$ function matrices A, B, C, D, P, Q, R, S occurring in* (13) *and* (17) *belong to the class H_2 and satisfy $A(0)P(0) = S(0)D(0) = I$. In this situation, the matrix L is J-unitary. Moreover, the solution (K, L) is unique except for substitutions of the form $K \to K(U + V)$, $L \to L(U \dotplus V)$, where U and V are constant unitary matrices. In addition, setting $\Psi = BA^{-1} = D^{-1}C$ and $\Phi = P^{-1}Q = RS^{-1}$ yields the generalized Schur representation of Ω, which is unique within the normalization just indicated.*

Proof. The first step consists in establishing the existence and uniqueness of (19). Consider both equations $\tilde{A} - \Omega C = P$ and $C - \tilde{\Omega}\tilde{A} = -e^{-i\theta}\tilde{R}$, which are part of $\Delta L = K$, subject to the constraints $P \in L_2^+$, $Q \in L_2^+$ and $P(0) = A(0)^{-1}$. In view of (15), they immediately yield the left equations in (12). Eliminating C produces $(I - \Gamma\Gamma^*)\tilde{A} = A(0)^{-1}$. The general solution is given by $A(e^{i\theta}) = M\tilde{Y}(e^{i\theta})$, where $Y \in L_2^-$ is uniquely determined from $(I - \Gamma\Gamma^*)Y = I$ while M is any $p \times p$ matrix satisfying $\tilde{M}M = Y(0)^{-1}$. Next, the remaining part of $\Delta L = K$ leads to the right equations (12), for which the existence of a "unique" solution is proved similarly.

Then, let us check that the matrix L is J-unitary. From $\Delta L = K$ one deduces, by straightforward computation,

$$\tilde{L}JL = \begin{bmatrix} AP + e^{i\theta}RC & e^{i\theta}(RD - AQ) \\ BP - SC & -(SD + e^{i\theta}BQ) \end{bmatrix}. \tag{20}$$

By construction, the right member of (20) belongs to the class L_1^+, whereas the left member is Hermitian. Hence the matrix (20) must be a constant and, in view of $A(0)P(0) = S(0)D(0) = I$, this constant must be J, which proves the claim.

Let us now derive further properties of K and L. In view of $\tilde{L}JL = J$, the equation $P = \tilde{A} - \Omega C$ yields $AP + \tilde{P}\tilde{A} - \tilde{P}P = I + \tilde{C}(I - \tilde{\Omega}\Omega)C$. As a result, the Hermitian part of the H_1-function $F(z) = A(z)P(z)$ is positive definite on the unit circle, so that $F(z)$ belongs to the class C. Together with $F(0) = I$, this property yields $F(z)^{-1} \in C$, implying that $F(z)^{-1}$, $A(z)^{-1}$ and $P(z)^{-1}$ are outer Hardy functions. A similar argument, based on $S = \tilde{D} - B\Omega$, leads to the conclusion that $D(z)^{-1}$ and $S(z)^{-1}$ are outer Hardy functions.

In agreement with the identity $CA = DB$ resulting from $LJ\tilde{L} = J$, define $\Psi(z) = B(z)A(z)^{-1} = D(z)^{-1}C(z)$. The preceding argument shows that $\Psi(z)$ is a Hardy function. On the other hand, $LJ\tilde{L} = J$ gives $I - \Psi\tilde{\Psi} = D^{-1}\tilde{D}^{-1}$ and $I - \tilde{\Psi}\Psi = \tilde{A}^{-1}A^{-1}$ on the unit circle. Hence $\Psi(z)$, $D^{-1}(z)$ and $A^{-1}(z)$ belong to the class S. Next, from (19) one obtains $KJ\tilde{K} = \Delta J \Delta$, which implies $QS = PR$ and allows one to put $\Phi(z) = P(z)^{-1}Q(z) = R(z)S(z)^{-1}$. In view of the preceding argument, $\Phi(z)$ is a Hardy function, which necessarily belongs to the class S as a consequence of the identities $I - \Phi\tilde{\Phi} = P^{-1}(I - \Omega\tilde{\Omega})\tilde{P}^{-1}$ and $I - \tilde{\Phi}\Phi = \tilde{S}^{-1}(I - \tilde{\Omega}\Omega)S^{-1}$ resulting from $KJ\tilde{K} = \Delta J \Delta$.

Finally, (19) implies (16) together with $\Omega = \tilde{A}(\tilde{\Psi} + e^{i\theta}\Phi)S$, which immediately gives the desired representation (11) with the required spectral factorization formulas (14). Since one has seen that, conversely, any generalized Schur representation produces an adequate factorization (19), this completes the proof of the theorem. □

3.2. Properties of the shift operation. Given a $p \times p$ function matrix $\Omega(e^{i\theta})$ satisfying (10), define Ω_k to be the k-shift of Ω, i.e., the function

$$\Omega_k(e^{i\theta}) = e^{ik\theta}\Omega(e^{i\theta}), \tag{21}$$

for any $k \in \mathbb{Z}$. Let Γ_k denote the block-Hankel operator associated with Ω_k. Since $\|\Gamma_k\| \leqq \|\Gamma_{k-1}\|$, the condition $\|\Gamma\| < 1$, which is assumed throughout, implies $\|\Gamma_k\| < 1$ for all $k > -q$, where q is either infinity or a positive integer. In this section it is shown that, in case $k \geqq 1$, the generalized Schur representation of Ω_k can be deduced from that of Ω by simple algebraic manipulations involving the k first Schur parameters of $\Psi(z)$. A similar result is given for the case $-q < k \leqq -1$, involving the parameters of $\Phi(z)$ instead of those of $\Psi(z)$. (The machinery appears clearly in the simple situation of § 2.1.)

Let $(E_0, E_{-1}, E_{-2}, \cdots)$ and $(\tilde{E}_1, \tilde{E}_2, \tilde{E}_3, \cdots)$ denote the sequences of Schur parameter matrices associated with the functions $\Psi(z)$ and $\Phi(z)$, respectively, where (Φ, Ψ) is the Schur pair occurring in the representation (11). Note that the first sequence is infinite and satisfies $\sum_{s=0}^{\infty} \|E_{-s}\|^2 < \infty$. (This follows from (14); see [5] and [6].) As pointed out below, the length of the second sequence is at least $q-1$. Next, for any $k \geqq 1$, define the class S function matrices $\Psi_k(z)$ and $\Phi_{-k}(z)$ by recursive application of the homographic transformations (cf. § 2.1)

$$z\Psi_k(z) = H(-E_{1-k})T(z)^{-1}[z\Psi_{k-1}(z)], \tag{22}$$

$$z\Phi_{-k}(z) = T(z)H(\tilde{E}_{1-k})[z\Phi_{1-k}(z)], \tag{23}$$

with the initialization $\Psi_0 = \Psi$, $\Phi_0 = \Phi$. Thus both families (Ψ_k) and (Φ_{-k}) satisfy the recurrence relation (1). A clear interpretation of (22) and (23) is obtained from considering the sequences of Schur parameters associated with Ψ_k and Φ_{-k}, namely

$$\begin{aligned} &\Psi_k(z)\colon (E_{-k}, E_{-1-k}, E_{-2-k}, \cdots), \\ &\Phi_{-k}(z)\colon (\tilde{E}_{1-k}, \tilde{E}_{2-k}, \tilde{E}_{3-k}, \cdots). \end{aligned} \tag{24}$$

On the other hand, starting from the canonical factorization (19), let us construct the $2p \times 2p$ function matrices $L_k(e^{i\theta})$ and $K_{-k}(e^{i\theta})$, for $k \geqq 1$, by means of the parameters $E_0, E_{-1}, \cdots, E_{1-k}$ of $\Psi(z)$, as follows:

$$L_k = T^k L T \prod_{s=0}^{k-1} [T^{-1}H(-\tilde{E}_{-s})]T^{-1}, \tag{25}$$

$$K_{-k} = T^k K T \prod_{s=0}^{k-1} [T^{-1}H(-\tilde{E}_{-s})]T^{-1}, \tag{26}$$

with $T = T(e^{i\theta})$. Here and in the sequel $\prod_{s=0}^{m} X_s$ means $X_0 X_1 \cdots X_m$ while $\prod_{s=m}^{0} X_s$ means $X_m X_{m-1} \cdots X_0$. From (25) and (26) let us now define $p \times p$ function matrices $A_k(e^{i\theta}), \cdots, D_k(e^{i\theta})$ and $P_{-k}(e^{i\theta}), \cdots, S_{-k}(e^{i\theta})$, by writing

$$L_k = \begin{bmatrix} \tilde{A}_k & \tilde{B}_k \\ C_k & D_k \end{bmatrix}, \qquad K_{-k} = \begin{bmatrix} P_{-k} & -e^{i\theta}Q_{-k} \\ -e^{-i\theta}\tilde{R}_{-k} & \tilde{S}_{-k} \end{bmatrix}. \tag{27}$$

THEOREM 2. *For any $k \geqq 1$, the canonical factorization* (19) *of the matrix $\Delta_k = T^k \Delta T^{-k}$ associated with Ω_k is determined from* (25) *and* (26) *to be $\Delta_k = K_{-k}L_k^{-1}$, so that the corresponding Schur pair is the pair (Φ_{-k}, Ψ_k) as given by* (22) *and* (23).

Proof. It suffices to consider the case $k = 1$ and then proceed by induction. From (25) and $E_0 = \Psi(0)$ it follows that A_1, B_1, C_1 and D_1 belong to the class H_2, while (26) directly implies $P_{-1}, Q_{-1}, R_{-1}, S_{-1} \in H_2$. In addition, $A_1(0)P_{-1}(0) = S_{-1}(0)D_1(0) = I$ appears as a consequence of $A(0)P(0) = S(0)D(0) = I$. Next, (19), (25) and (26) yield $\Delta_1 = T\Delta T^{-1} = TKL^{-1}T^{-1} = K_{-1}L_1^{-1}$. By Theorem 1, this proves the first assertion. (Note that the J-unitarity of L_1 is immediate from (25).)

To establish the second part let us put the functions $\Psi' = B_1A_1^{-1}$ and $\Phi' = P_{-1}^{-1}Q_{-1}$. It follows from Theorem 1 that (Φ', Ψ') is the Schur pair of Ω_1. On the other hand, it is easily checked, by use of (25) and (26), that $z\Psi'$ and $z\Phi'$ coincide with the right members of (22) and (23), respectively. This concludes the proof. □

An alternative version of (25) will be useful in the sequel. Defining the permutation matrix W as in (2) one can verify, by elementary computation, that (25) is equivalent to

$$WL_kW = T^{-k}(WLW)T^{-1}\prod_{s=0}^{k-1}[TH(-E_{-s})]T. \tag{28}$$

COROLLARY 3. *Assume the block-Hankel operator Γ_{-k} corresponding to Ω_{-k} satisfies $\|\Gamma_{-k}\| < 1$, for a given positive integer k. Then the length of the Schur sequence of $\Phi(z)$ is at least equal to k, and the canonical factorization of the matrix $\Delta_{-k} = T^{-k}\Delta T^k$ is given by $\Delta_{-k} = K_kL_{-k}^{-1}$, with*

$$L_{-k} = T^{-k}LT\prod_{s=1}^{k}[H(\tilde{E}_s)T]T^{-1}, \tag{29}$$

$$K_k = T^{-k}KT\prod_{s=1}^{k}[H(\tilde{E}_s)T]T^{-1}. \tag{30}$$

As a consequence, the functions Φ_k and Ψ_{-k} in the corresponding Schur pair are characterized in terms of their sequences of Schur parameters as in (24), *with k replaced by $-k$.*

Proof. It suffices to apply Theorem 2 with Ω_{-k} substituted for Ω and then to interpret the result backwards. The details are omitted. □

Let $M_k(e^{i\theta})$ and $N_k(e^{i\theta})$ denote the Blaschke–Potapov products occurring in (25) and (28), respectively; i.e.,

$$M_k = \prod_{s=k-1}^{0}[H(-\tilde{E}_{-s})T], \qquad N_k = \prod_{s=0}^{k-1}[TH(-E_{-s})]. \tag{31}$$

Thus $M_k = e^{ik\theta}W\tilde{N}_kW$. For application to convergence problems it will prove useful to express M_k and N_k in terms of Szegö orthogonal polynomial matrices [5]. Let $F^+(z)$ and $F^-(z)$ be the class C function matrices associated with $\Psi(z)$ and $-\Psi(z)$, respectively; i.e.,

$$F^{\pm}(z) = [I \mp z\Psi(z)][I \pm z\Psi(z)]^{-1}. \tag{32}$$

Define $X_k^{\pm}(z)$ and $Y_k^{\pm}(z)$ to be, respectively, the left and right orthogonal polynomial matrices of degree k associated with (32). Applying the results of [6] one obtains

$$N_k = \frac{1}{2}\begin{bmatrix} Y_k^- + Y_k^+ & \hat{X}_k^- - \hat{X}_k^+ \\ Y_k^- - Y_k^+ & \hat{X}_k^- + \hat{X}_k^+ \end{bmatrix}, \tag{33}$$

with $\hat{G}_k(z) = z^k\tilde{G}_k(1/\bar{z})$, denoting the reciprocal of a polynomial matrix $G_k(z)$ of degree k. An expression quite similar to (33) holds for $M_k = e^{ik\theta}W\tilde{N}_kW$. As a result,

using (25) and (28), one deduces

$$
\begin{aligned}
2\tilde{A}_k &= (\tilde{A}+e^{-i\theta}\tilde{B})\hat{X}_k^- + (\tilde{A}-e^{-i\theta}\tilde{B})\hat{X}_k^+,\\
2e^{i(k+1)\theta}C_k &= (D+e^{i\theta}C)\hat{X}_k^- - (D-e^{i\theta}C)\hat{X}_k^+,\\
2\tilde{D}_k &= \hat{Y}_k^-(\tilde{D}+e^{-i\theta}\tilde{C}) + \hat{Y}_k^+(\tilde{D}-e^{-i\theta}\tilde{C}),\\
2e^{i(k+1)\theta}B_k &= \hat{Y}_k^-(A+e^{i\theta}B) - \hat{Y}_k^+(A-e^{i\theta}B).
\end{aligned}
\tag{34}
$$

Note that the expressions under brackets in the right members of (34) are the inverses of the spectral factors of the Hermitian part of $F^{\pm}(e^{i\theta})$. Indeed, it is easily seen that one has

$$
\begin{aligned}
\operatorname{Herm} F^{\pm} &= (D\pm e^{i\theta}C)^{-1}(\tilde{D}\pm e^{-i\theta}\tilde{C})^{-1}\\
&= (\tilde{A}\pm e^{-i\theta}\tilde{B})^{-1}(A\pm e^{i\theta}B)^{-1}.
\end{aligned}
\tag{35}
$$

3.3. Dual representation. It turns out that interchanging the roles of $\Phi(z)$ and $\Psi(z)$ in (11) exhibits an important duality in the theory. However, this makes sense only for a restricted set of functions $\Omega(e^{i\theta})$, because Φ does not generally enjoy all properties required from Ψ. (In particular, (14) may fail to exist when Ψ is replaced by Φ.) In fact, an appropriate assumption in the present context is $\|\Omega\|_\infty < 1$, i.e., strict inequality in (10), which obviously implies $\|\Gamma\| < 1$. In this case, there exist spectral factorizations

$$
(I-\Omega\tilde{\Omega})^{-1} = \tilde{G}G, \qquad (I-\tilde{\Omega}\Omega)^{-1} = H\tilde{H},
\tag{36}
$$

where $G(z)$, $H(z)$, $G(z)^{-1}$ and $H(z)^{-1}$ are $p\times p$ function matrices belonging to the Hardy class H_∞. Note that G and H are uniquely determined within a left and a right unitary factor, respectively.

Let us then define eight matrix functions $A'(z)$, $B'(z)$, $C'(z)$, $D'(z)$, $P'(z)$, $Q'(z)$, $R'(z)$ and $S'(z)$, all of class H_2, as follows:

$$
\begin{aligned}
&A' = SH, \quad B' = RH, \quad C' = GQ, \quad D' = GP,\\
&P' = H^{-1}D, \quad Q' = H^{-1}C, \quad R' = BG^{-1}, \quad S' = AG^{-1},
\end{aligned}
\tag{37}
$$

where A, B, C, D, P, Q, R, S yield the canonical factorization (19) of Δ, while G and H are determined from (36). Next, we construct the $2p\times 2p$ matrices L' and K' as in (13) and (17), with $A', \cdots, S'$ substituted for $A, \cdots, S$. Direct computation from (37) gives

$$
L' = F(TJW)K(WJT^{-1}), \qquad K' = \tilde{F}^{-1}(TJW)L(WJT^{-1}),
\tag{38}
$$

with $F = \tilde{H} \pm G$, where W, J and $T = T(e^{i\theta})$ are as in (2). From $KJ\tilde{K} = \Delta J\Delta = JW(\tilde{F}F)^{-1}W$, one readily deduces $L'J\tilde{L}' = J$. On the other hand, the canonical factorization (19), written in the form $\Delta = (JLJ)\tilde{K}$, leads to

$$
\tilde{F}^{-1}(TW\Delta WT^{-1})\tilde{F} = K'(J\tilde{L}'J),
\tag{39}
$$

by straightforward computation from (38). Let Δ' denote the left member of (39). By definition, Δ' appears as the matrix (18) where Ω is replaced by the function Ω' given by

$$
\Omega'(e^{i\theta}) = e^{i\theta}H(e^{i\theta})^{-1}\tilde{\Omega}(e^{i\theta})\tilde{G}(e^{i\theta}).
\tag{40}
$$

Note indeed that $\tilde{\Omega}' = e^{-i\theta}\tilde{G}^{-1}\Omega H$ in view of (36). A further consequence of (36) and (40) is $\|\Omega'(e^{i\theta})\| = \|\Omega(e^{i\theta})\|$. In the sequel, Ω' is referred to as the *dual* of Ω.

Using the identities above and applying Theorem 1 we now immediately obtain the following result about duality.

THEOREM 4. *The canonical factorization* $\Delta' = K'(L')^{-1}$ *relative to the dual function* Ω' *is deduced from that of* Δ *by* (38); *the Schur pair* $(\Phi'\ \Psi')$ *of* Ω' *is given by* $\Phi' = \Psi$, $\Psi' = \Phi$.

It is interesting to note that the spectral factors G' and H' associated with Ω', as in (36), are given, up to normalization, by $G'(z) = H(z)$ and $H'(z) = G(z)$. Let us now derive some matrix inequalities that will play an important role in the next section:

$$I \leqq \tilde{A}(0)A(0) \leqq \tilde{G}(0)G(0), \qquad I \leqq D(0)\tilde{D}(0) \leqq H(0)\tilde{H}(0). \tag{41}$$

The argument goes as follows. From the fact that $A(z)^{-1}$ belongs to the class S one deduces $\tilde{A}(0)A(0) \geqq I$ and $A(0)\tilde{A}(0) \geqq I$. Similar results hold for A', D and D'. Next, $P(0)A(0) = I$ yields $D'(0)A(0) = G(0)$, by (37); hence $\tilde{A}(0)A(0) \leqq \tilde{G}(0)G(0)$ in view of $\tilde{D}'(0)D'(0) \geqq I$. The result $D(0)\tilde{D}(0) \leqq H(0)\tilde{H}(0)$ is proved similarly.

Let us finally examine the interplay between duality and shifting. Note that, in terms of the Schur parameters, duality is expressed by $E'_k = \tilde{E}_{1-k}$ (see Theorem 4). In agreement with this, it appears that shifting Ω to the left can be interpreted as shifting Ω' to the right. More precisely, in view of the fact that the spectral factorizations (36) are not affected by shifting, the k-shift $\Omega'_k = e^{ik\theta}\Omega'$ is the dual of $\Omega_{-k} = e^{-ik\theta}\Omega$ for any $k \in \mathbb{Z}$. Thus, according to Theorem 4, the J-unitary matrix L'_k occurring in the canonical factorization relative to Ω'_k is determined by

$$A'_k = S_k H, \quad B'_k = R_k H, \quad C'_k = GQ_k, \quad D'_k = GP_k. \tag{42}$$

4. Asymptotic behavior. The present section is devoted to the question of the behavior of the generalized Schur representation of the shifted function $\Omega_k(e^{i\theta})$ when k tends to $+\infty$ or to $-\infty$. This question actually plays an important role in the analysis of half-plane block-Toeplitz systems [9]. (See [8] for the scalar case.)

4.1. Convergence in the mean. Let us first state a lemma which turns out to be the main tool for treating L_2-convergence. For any $k \in \mathbb{Z}$ define the $p \times p$ function matrices $X_k(e^{i\theta})$ and $Y_k(e^{i\theta})$ of class L_2 as follows:

$$X_k(e^{i\theta}) = e^{i(1-k)\theta}Q_{-k}(e^{i\theta})\tilde{D}_k(0), \tag{43}$$

$$Y_k(e^{i\theta}) = e^{i(1-k)\theta}\tilde{A}_k(0)R_{-k}(e^{i\theta}). \tag{44}$$

LEMMA 5. *Assume* $\|\Omega\|_\infty \leqq 1$. *For all integers* k *and* n *with* $k \geqq n$ *one has both matrix inequalities*

$$\frac{1}{2\pi}\int_0^{2\pi} (\tilde{X}_k - \tilde{X}_n)(X_k - X_n)\, d\theta \leqq D_n(0)\tilde{D}_n(0) - D_k(0)\tilde{D}_k(0), \tag{45}$$

$$\frac{1}{2\pi}\int_0^{2\pi} (Y_k - Y_n)(\tilde{Y}_k - \tilde{Y}_n)\, d\theta \leqq \tilde{A}_n(0)A_n(0) - \tilde{A}_k(0)A_k(0). \tag{46}$$

Proof. By definition, $X_k = (\Omega D_k - e^{-ik\theta}\tilde{B}_k)\tilde{D}_k(0)$. Hence, applying the obvious inequality $(\tilde{M}\tilde{\Omega} + \tilde{N})(\Omega M + N) \leqq \tilde{M}M + \tilde{N}N + \tilde{M}\tilde{\Omega}N + \tilde{N}\Omega M$, one can write

$$(\tilde{X}_k - \tilde{X}_n)(X_k - X_n) \leqq F_{k,k} + F_{n,n} - F_{k,n} - F_{n,k}, \tag{47}$$

where $F_{k,n}(=\tilde{F}_{n,k})$ is given by

$$\begin{aligned} F_{k,n} &= D_k(0)[\tilde{D}_k(D_n - \tilde{\Omega}_n\tilde{B}_n) + e^{i(k-n)\theta}B_k(\tilde{B}_n - \Omega_n D_n)]\tilde{D}_n(0) \\ &= D_k(0)[\tilde{D}_k\tilde{S}_{-n} - e^{i(k-n+1)\theta}B_k Q_{-n}]\tilde{D}_n(0). \end{aligned} \tag{48}$$

In view of $D_n(0)S_{-n}(0) = I$, integration of (48) yields

$$\frac{1}{2\pi}\int_0^{2\pi} F_{k,n}(e^{i\theta})\,d\theta = D_k(0)\tilde{D}_k(0) \quad \text{for } k \geqq n. \tag{49}$$

The desired result (45) directly follows from (47) and (49). The proof of (46) is similar and left to the reader. □

THEOREM 6. *Assume* $\|\Omega\|_\infty < 1$. *The matrices* $A_k(e^{i\theta})$, $B_k(e^{i\theta})$, $C_k(e^{i\theta})$, $D_k(e^{i\theta})$ *occurring in the generalized Schur representation of* $\Omega_k(e^{i\theta})$ *have the following* L_2*-convergence properties*:

$$\text{l.i.m.}_{k\to\infty} \tilde{A}_k(0)A_k = I, \qquad \text{l.i.m.}_{k\to-\infty} \tilde{A}_k(0)A_k = \tilde{G}(0)G,$$

$$\text{l.i.m.}_{k\to\infty} B_k = 0, \qquad \text{l.i.m.}_{k\to-\infty} e^{ik\theta}D_k(0)B_k = H(0)\tilde{H}\tilde{\Omega},$$

$$\text{l.i.m.}_{k\to\infty} C_k = 0, \qquad \text{l.i.m.}_{k\to-\infty} e^{ik\theta}C_kA_k(0) = \tilde{\Omega}\tilde{G}G(0),$$

$$\text{l.i.m.}_{k\to\infty} D_k\tilde{D}_k(0) = I, \qquad \text{l.i.m.}_{k\to-\infty} D_k\tilde{D}_k(0) = H\tilde{H}(0).$$

Proof. In view of (41) and (45), the doubly infinite sequence of matrices $D_k(0)\tilde{D}_k(0)$ is bounded and monotone; hence it converges for $k \to \pm\infty$. As a result it appears from (45), owing to the Cauchy criterion, that the sequence of $X_k(e^{i\theta})$ converges in the L_2-sense, for $k \to \pm\infty$, to a well-defined function matrix $X_\pm(e^{i\theta})$. Now, since Q_{-k} belongs to L_2^+, it is clear from (43) that X_- has to be the zero function. Thus, given any positive real number ε, one can write

$$\frac{1}{2\pi}\int_0^{2\pi} \tilde{X}_{-k}(e^{i\theta})X_{-k}(e^{i\theta})\,d\theta \leqq \varepsilon I, \tag{50}$$

provided $k \geqq k_0(\varepsilon)$. Let us put $c = 1/(1-\|\Omega\|_\infty^2)$. Then (36) yields $\tilde{G}(e^{i\theta})G(e^{i\theta}) \leqq cI$. Hence, using (41), (42) and (43), one deduces from (50)

$$\begin{aligned}\frac{1}{2\pi}\int \tilde{C}'_kC'_k\,d\theta &= \frac{1}{2\pi}\int \tilde{Q}_k\tilde{G}GQ_k\,d\theta \\ &\leqq \frac{c}{2\pi}\int D_{-k}(0)^{-1}\tilde{X}_{-k}X_{-k}\tilde{D}_{-k}(0)^{-1}\,d\theta \\ &\leqq c\varepsilon[\tilde{D}_{-k}(0)D_{-k}(0)]^{-1} \leqq c\varepsilon I,\end{aligned} \tag{51}$$

which yields l.i.m. $C'_k = 0$ for $k \to \infty$. Next, apply $A'\tilde{A}' - \tilde{C}'C' = I$, resulting from the J-unitarity of L'. As a consequence of (51) and (41) one readily deduces

$$\frac{1}{2\pi}\int [I - \tilde{A}'_k(0)A'_k][I - \tilde{A}'_kA'_k(0)]\,d\theta \leqq c\varepsilon\tilde{H}(0)H(0), \tag{52}$$

hence l.i.m. $\tilde{A}'_k(0)A'_k = I$, for $k \to \infty$. A similar argument, based on (46) instead of (45), leads to $B'_k \to 0$ and $D'_k\tilde{D}'_k(0) \to I$. Thus the dual versions of the desired results are established in the case $k \to +\infty$. The primal versions follow of course from the same argument.

Let us now consider the case $k \to -\infty$. Define the block-diagonal matrix $M_k = \tilde{D}_k(0) \dotplus A_k(0)$. In view of $A_k(0) = D'_{-k}(0)^{-1}G(0)$ and $D_k(0) = H(0)A'_{-k}(0)^{-1}$, the results of the first part imply $L'_{-k}M_k \to F(0)$, in the L_2-sense, with $F = \tilde{H} + G$ as above.

Then, expressing L' in terms of K as in (38), one obtains $WK_{-k}WM_k \to F^{-1}F(0)$, hence

$$\text{l.i.m.}_{k\to-\infty} \Delta(T^{-k}L_kT^k)WM_k = WF^{-1}F(0), \tag{53}$$

by use of $\Delta_k L_k = K_{-k}$ with $\Delta_k = T^k \Delta T^{-k}$. Finally, multiplying (53) to the left by ΔJ and using $\Delta J \Delta = JW(\tilde{F}F)^{-1}W$ together with (36) gives

$$\text{l.i.m.}_{k\to-\infty} (T^{-k}L_kT^k)WM_k = (J\Delta J)W\tilde{F}F(0). \tag{54}$$

In compact form, this is precisely the desired result. Hence the theorem is proved. □

As a direct consequence of Theorem 6, it appears that the bounds (41) are achieved in the limits

$$\begin{aligned} &\lim_{k\to+\infty} \tilde{A}_k(0)A_k(0) = I, \qquad \lim_{k\to-\infty} \tilde{A}_k(0)A_k(0) = \tilde{G}(0)G(0), \\ &\lim_{k\to+\infty} D_k(0)\tilde{D}_k(0) = I, \qquad \lim_{k\to-\infty} D_k(0)\tilde{D}_k(0) = H(0)\tilde{H}(0). \end{aligned} \tag{55}$$

COROLLARY 7. *Let the spectral factorizations* (36) *and the representation of* Ω_k *be normalized in such a way that* $G(0)$, $H(0)$, $A_k(0)$ *and* $D_k(0)$ *are Hermitian positive definite. Then one has the* L_2*-convergence properties* $A_k \to I$, $B_k \to 0$, $C_k \to 0$, $D_k \to I$ *for* $k \to +\infty$ *and* $A_k \to G$, $e^{ik\theta}B_k \to \tilde{H}\tilde{\Omega}$, $e^{ik\theta}C_k \to \tilde{\Omega}\tilde{G}$, $D_k \to H$ for $k \to -\infty$.

Proof. This immediately follows from Theorem 6, since (55) yields $A_k(0) \to I$, $D_k(0) \to I$ for $k \to \infty$ and $A_k(0) \to G(0)$, $D_k(0) \to H(0)$ for $k \to -\infty$. □

COROLLARY 8. *Under the same conditions as in Corollary* 7, *the Schur pairs* (Φ_{-k}, Ψ_k) *of the shifted functions* Ω_k *satisfy*

$$\begin{aligned} &\text{l.i.m.}_{k\to+\infty} \Psi_k = 0, \qquad \text{l.i.m.}_{k\to-\infty} e^{ik\theta}\Psi_k = e^{-i\theta}\Omega', \\ &\text{l.i.m.}_{k\to+\infty} \Phi_k = 0, \qquad \text{l.i.m.}_{k\to-\infty} e^{ik\theta}\Phi_k = e^{-i\theta}\Omega. \end{aligned}$$

Proof. Since A_k^{-1} belongs to the class S, the property $B_k \to 0$ in Theorem 6 implies $\Psi_k \to 0$ for $k \to \infty$. The second result concerning Ψ_k follows from the identity

$$e^{-i\theta}\Omega' - e^{ik\theta}\Psi_k = (\tilde{H}\tilde{\Omega} - e^{ik\theta}B_k)G^{-1} + e^{ik\theta}\Psi_k(A_k - G)G^{-1}. \tag{56}$$

Indeed, in view of $G^{-1} \in H_\infty$ and $\Psi_k \in S$, it appears from Corollary 7 that the right member of (56) tends to zero for $k \to -\infty$. The desired results concerning Φ_k are proved in a similar manner. □

4.2. Convergence in Banach algebras. The developments below are mainly based on formulas (34). We shall use matrix versions of remarkable theorems due to Baxter [3] about convergence of Szegö's orthogonal polynomials; the appropriate generalization to the matrix case has been recently made by Geronimo [11]. Let α be an even real-valued function defined over the integers, satisfying $\alpha(m) \geqq 1$, $\alpha(m+n) \leqq \alpha(m)\alpha(n)$ for all $m, n \in \mathbb{Z}$, and $\lim \alpha(m)^{1/m} = 1$ for $m \to \infty$. Given an integrable function matrix $X(e^{i\theta}) \sim \sum X_m e^{im\theta}$, define

$$|X|_\alpha = \sum_{m=-\infty}^{+\infty} \alpha(m)\|X_m\|. \tag{57}$$

The set $\mathcal{B}_\alpha$ consisting of the functions X with finite value of (57) clearly is a Banach algebra for the norm $|X|_\alpha$. (The particular choice $\alpha(m) = 1$ for all m leads to the classical Wiener algebra.)

Henceforth let us assume that the function Ω belongs to $\mathcal{B}_\alpha$, for a given admissible α. Using a general theorem concerning compact operators in Banach spaces with two norms [1], one deduces that the function matrices A, B, C, D, as defined from (12), belong to the subalgebra $\mathcal{B}_\alpha^+$ of $\mathcal{B}_\alpha$. As a consequence, in view of the Wiener–Lévy theorem, one has $F^\pm(e^{i\theta}) \in \mathcal{B}_\alpha^+$ (see (32)). Then Baxter's theorem [3], [11] together with the results of [6] imply that the sequences of polynomial matrices $\hat{X}_k^\pm$ and $\hat{Y}_k^\pm$ converge, in the sense of the norm (57), to the inverse of the right and left spectral factors of Herm $F^\pm$, respectively. Thus, according to (35), one has

$$|\hat{X}_k^\pm - (A \pm e^{i\theta}B)|_\alpha \to 0, \qquad |\hat{Y}_k^\pm - (D \pm e^{i\theta}C)|_\alpha \to 0 \tag{58}$$

when $k \to \infty$, for a suitable normalization of A and D. It is interesting to observe that (58) can be derived from the condition $\sum_{s=0}^{\infty} \alpha(s)\|E_{-s}\| < \infty$, which actually is equivalent to $A, B, C, D \in \mathcal{B}_\alpha^+$ (see [3], [11]).

In case $\|\Omega\|_\infty < 1$ the assumption $\Omega \in \mathcal{B}_\alpha$ implies $G, H, G^{-1}, H^{-1} \in \mathcal{B}_\alpha^+$ (cf. [12]). Hence the properties mentioned above hold true for the dual function matrix Ω'. The main results concerning the generalized Schur representation in the framework of the Banach algebra $\mathcal{B}_\alpha$ are collected in the following theorem.

THEOREM 9. *Let $\|\Omega\|_\infty < 1$. Then the conditions $\Omega \in \mathcal{B}_\alpha$ and $\sum_{m=-\infty}^{\infty} \alpha(m)\|E_m\| < \infty$ are equivalent. Either of them implies $A_k, B_k, C_k, D_k \in \mathcal{B}_\alpha^+$ for all k, with the convergence properties $A_k \to I$, $e^{ik\theta}B_k \to 0$, $e^{ik\theta}C_k \to 0$, $D_k \to I$ when $k \to +\infty$ and $A_k \to G$, $e^{ik\theta}B_k \to \tilde{H}\tilde{\Omega}$, $e^{ik\theta}C_k \to \tilde{\Omega}\tilde{G}$, $D_k \to H$ when $k \to -\infty$, in the sense of the norm $|\cdot|_\alpha$.*

Proof. This essentially follows from applying Baxter's techniques [3], [11] to the results of [6]. In particular, the first set of convergence properties (for $k \to +\infty$) directly follows from (34), (58) and $LJ\tilde{L} = J$. By duality, A_k', B_k', C_k', D_k' enjoy similar properties. Interpreting these in terms of the functions A_{-k}, B_{-k}, C_{-k}, D_{-k} as in the second part of the proof of Theorem 6, one obtains the desired properties for $k \to -\infty$. The details are not repeated here. □

Let us make a final remark which is important from the application viewpoint [8]. If $\Omega(e^{i\theta})$ is an infinitely differentiable periodic function of θ, then Ω belongs to the algebra $\mathcal{B}_\alpha$ for all α with polynomial growth. In this case the properties quoted in Theorem 9 are very strong. For example, the statements $A_k \in \mathcal{B}_\alpha$ and $A_k \to G$ mean that $A_k(e^{i\theta})$ is infinitely differentiable and that its nth derivative converges uniformly to the nth derivative of $G(e^{i\theta})$ for any given $n \geqq 0$.

Acknowledgments. The authors are most grateful to C. Foias for calling their attention to the relationship between their work and certain results by Adamjan, Arov and Krein. Thanks are also due to J. S. Geronimo for communicating his results.

REFERENCES

[1] V. M. ADAMJAN, D. Z. AROV AND M. G. KREIN, *Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem*, Math. USSR-Sb., 15 (1971), pp. 31–73.

[2] ———, *Infinite Hankel block matrices and related extension problems*, Amer. Math. Soc. Transl., 111 (1978), pp. 133–156.

[3] G. BAXTER, *A convergence equivalence related to polynomials orthogonal on the unit circle*, Trans. Amer. Math. Soc., 99 (1961), pp. 471–487.

[4] H. CHANG AND J. K. AGGARWAL, *Design of two-dimensional semicausal recursive filters*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 1051–1059.

[5] P. DELSARTE, Y. GENIN AND Y. KAMP, *Orthogonal polynomial matrices on the unit circle*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 149–160.

[6] ———, *Schur parametrization of positive definite block-Toeplitz systems*, SIAM J. Appl. Math., 36 (1979), pp. 34–46.

[7] ———, *Generalized Schur parametrization*, Proc. 4th Int. Symp. Mathematical Theory of Networks and Systems, Delft, the Netherlands, July 1979, Western Periodicals, North Hollywood, CA, pp. 290–296.

[8] ———, *Half-plane Toeplitz systems*, IEEE Trans. Information Theory, IT-25 (1980), pp. 465–474.

[9] ———, *Half-plane minimization of matrix-valued quadratic functionals*, this Journal, this issue, pp. 192-211.

[10] M. P. EKSTROM AND J. W. WOODS, *Two-dimensional spectral factorization with applications in recursive digital filtering*, IEEE Trans. Acoust. Speech Signal Process., ASSP-24 (1976), pp. 115–128.

[11] J. S. GERONIMO, *Matrix orthogonal polynomials on the unit circle*, to be published.

[12] I. C. GOHBERG AND I. A. FEL'DMAN, *Convolution Equations and Projection Methods for their Solution*, American Mathematical Society, Providence, RI, 1974.

[13] T. L. MARZETTA, *A linear prediction approach to two-dimensional spectral factorization and spectral estimation*, Ph. D. Dissertation, Dept. Electr. Eng. and Comp. Sci., MIT, Cambridge, MA, 1978.

[14] R. NEVANLINNA, *Über beschränkte analytische Funktionen*, Ann. Acad. Sci. Fenn., A15 (1920), pp. 1–75.

[15] B. T. O'CONNOR AND T. S. HUANG, *Stability of general two-dimensional recursive digital filters*, IEEE Trans. Acoust. Speech Signal Process., ASSP-26 (1978), pp. 550–560.

[16] G. PICK, *Über die Beschränkungen analytischer Funktionen, welche durch vorgegebene Funktionswerte bewirkt werden*, Math. Ann., 77 (1916), pp. 7–23.

[17] V. P. POTAPOV, *The multiplicative structure of J-contractive matrix functions*, Amer. Math. Soc. Transl., 15 (1960), pp. 131–243.

[18] J. SCHUR, *Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind*, Z. Reine Angew. Math., 147 (1917), pp. 205–232 and 148 (1918), pp. 122–145.

[19] T. TAKAGI, *On an algebraic problem related to an analytic theorem of Carathéodory and Fejér*, Japan J. Math., 1 (1924), pp. 83–93 and 2 (1925), pp. 13–17.

SIAM. J. ALG. DISC. METH.
Vol. 2, No. 2, June 1981

0196-5212/81/0202-0004 $01.00/0

WHITNEY CONNECTIVITY OF MATROIDS*

THOMAS INUKAI† AND LOUIS WEINBERG‡

Abstract. A new definition of matroid connectivity is introduced and its properties are investigated in this paper. Vertex connectivity of graphs is expressed in an algebraic form and generalized to matroids. This generalized connectivity is called the Whitney connectivity of matroids. It is shown that the Whitney connectivity of the polygon matroid of a graph is the same as the vertex connectivity of the graph provided the graph is connected. Various properties of Whitney matroid connectivity and comparison with Tutte connectivity are also examined.

1. Introduction. A matroid may be defined as a generalization of a graph obtained by abstracting certain topological properties of polygons and cut-sets of the graph, and accordingly many graph theorems have been generalized to matroids. However, in the process of abstraction, some important concepts in graph theory are lost in matroids; for example, there is no matroid concept corresponding to a vertex of a graph. Thus, although the vertex connectivity of graphs, a concept due to Whitney [8], is accepted by most researchers as the standard definition of graph connectivity, Tutte introduces another definition of graph connectivity [4] in such a manner as to allow its generalization to matroids [6]. He then justifies his matroid concept by showing that the connectivity of a connected graph is equal to the connectivity of its polygon matroid, and, furthermore, that the connectivity of a matroid and its dual is the same. Its usefulness was clearly demonstrated by Tutte's applying it to 3-connected matroids, where it led to the useful concept of the whirl matroid, which along with the wheel concept for 3-connected graphs yielded a satisfying theory for 3-connected matroids [6]. It is this theory that was used by the present authors in formulating the first efficient algorithm for determining whether a general matroid is realizable as a graph [3]. Tutte also used his matroid connectivity concept to establish Menger's theorem for matroids [5].

In this paper a new definition of matroid connectivity is proposed, and its basic properties are derived. Since the definition reduces, in the special case of graphs, to the vertex connectivity of a graph, we call this connectivity the Whitney connectivity of a matroid. As we shall see, we use the same connectivity function as Tutte and differ only in the second condition; that is, Tutte requires $\min(|S|, |\bar{S}|) \geqq n$ whereas we require $\min(r(\mathbf{M}\times S), r(\mathbf{M}\times \bar{S})) \geqq n$.

The contents of this paper are organized in the following fashion. After the introduction of some basic graph and matroid terminology, an algebraic form of Whitney graph connectivity is derived from the original definition. Whitney connectivity of matroids is then a straightforward generalization of the algebraic form, and this generalization is justified by proving that the generalized matroid connectivity of the polygon matroid of a connected graph coincides with the vertex connectivity of the graph. Various properties of Whitney matroid connectivity are investigated, and comparisons with the connectivity definition of Tutte are also presented. For example, we show that, contrary to what is true for Tutte connectivity, the Whitney connectivity of a matroid is not equal, in general, to that of its dual.

* Received by the editors January 26, 1978, and in final form November 4, 1980.

† COMSAT Laboratories, 22300 Comsat Drive, Clarksburg, Maryland 20734. Part of this work was performed at the City College of the City University of New York.

‡ City College and Graduate School, City University of New York, New York, New York 10031.

2. Definitions. Let $G=(V, E)$ be a graph, and let $S \subseteq E(G)$; G may contain multiple edges. The graph *reduction* and *contraction* operations are defined as usual and are denoted by $G \cdot S$ and $G \times S$, respectively.

A connected graph with each vertex having valence two is called a *polygon graph*, and $S \subseteq E(G)$ is a *polygon* of G if $G \cdot S$ is a polygon graph. A graph B is called a *cut-set graph* if $V(B)=\{v_1, v_2\}$, $E(B) \neq \varnothing$, and the ends of each member of $E(B)$ are v_1 and v_2. A subset T of $E(G)$ is a *cut-set* of G if $G \times T$ is a cut-set graph.

A graph G is said to be *n-connected* if the deletion of any $n-1$ or fewer vertices and their incident edges results in a connected graph, and the *connectivity* of G is n if G is n-connected but not $(n+1)$-connected. If there does not exist such an integer, then G has an arbitrarily high connectivity (denoted by ∞). Thus a complete graph K_p has ∞ connectivity. This definition of graph connectivity was first introduced by Whitney [8], and is generally termed the *vertex connectivity* of a graph.

The *rank* of G, $r(G)$, is the number of elements in a spanning forest of G. If $c(G)$ denotes the number of components of G, then $\mu(G)=|E|-r(G)=|E|-|V|+c(G)$ is called the *nullity* of G. Let S be a nonnull proper subset of $E(G)$. Then $\eta(G; S, \bar{S})$ denotes the number of common vertices of $G \cdot S$ and $G \cdot \bar{S}$, where $\bar{S}=E-S$.

Let $\mathbf{M}=(E, \mathbf{C})$ be a matroid which satisfies the circuit axioms. The members of E and $\mathbf{C}$ are referred to as *cells* and *circuits* of $\mathbf{M}$, respectively. The *rank* and *nullity* of $\mathbf{M}$ are denoted by $r(\mathbf{M})$ and $\mu(\mathbf{M})$, and no confusion should arise because of the use of the same symbols r and μ for matroid and graph invariants. $\mathbf{M}^*=(E, \mathbf{C}^*)$ denotes the *dual matroid* of $\mathbf{M}$, and the members of $\mathbf{C}^*$ are called *cocircuits* of $\mathbf{M}$.

Let $G=(V, E)$ be a graph, and let $\mathbf{C}_P$ and $\mathbf{C}_B$ be the classes of polygons and cut-sets of G, respectively. Then $\mathbf{P}(G)=(E, \mathbf{C}_P)$ and $\mathbf{B}(G)=(E, \mathbf{C}_B)$ satisfy the matroid axioms and are termed the *polygon* and *bond matroids* of $\mathbf{M}$. (Note that $\mathbf{P}^*(G)=\mathbf{B}(G)$.) A matroid $\mathbf{M}$ is called *graphic* or *cographic* if there is a graph G such that $\mathbf{M}=\mathbf{B}(G)$ or $\mathbf{P}(G)$, respectively.

Let $S \subseteq E$ and $\mathbf{C} \times S=\{C \mid C \in \mathbf{C}$ and $C \subseteq S\}$. Then $\mathbf{M} \times S=(S, \mathbf{C} \times S)$ is the *contraction* of $\mathbf{M}$ to S. If $\mathbf{C} \cdot S$ denotes the class of nonnull minimal intersections of the members of $\mathbf{C}$ with S, then $\mathbf{M} \cdot S=(S, \mathbf{C} \cdot S)$ is the *reduction* of $\mathbf{M}$ to S. In the above definitions of matroid contraction and reduction we are following Tutte, and do so throughout the paper. It should, however, be pointed out that these definitions are not the only ones used in the literature. There is also a definition that reverses the terms so that a matroid is said to be graphic if and only if it is a polygon matroid, and the reduction of a polygon matroid gives the polygon matroid of the reduced graph.

The reader may refer to Welsh [7] for other matroid terminology used in this paper.

3. Graph connectivity. In this section we formulate an equivalent definition of the vertex connectivity of graphs in a form that is more convenient for a generalization to matroids.

Let $G=(V, E)$ be connected. Then $\lambda(G)$ denotes the least integer n which satisfies the following conditions:

$$\eta(G; S, \bar{S})=n \quad \text{and} \quad \min(r(G \cdot S), r(G \cdot S)) \geqq n,$$

where S is a nonnull proper subset of E. If such an integer does not exist, we then write $\lambda(G)=\infty$. It shall be proved in the following that $\lambda(G)$ is equal to the vertex connectivity of G.

THEOREM 1. *Let G be a graph containing at least $n+1$ vertices. Then the vertex connectivity of G is $\lambda(G)$.*

This theorem is obtained as consequence of Lemmas 1 and 2 described below.

Let $G=(V, E)$ be a graph of connectivity n, where n is a positive finite integer. According to the definition of graph connectivity, we can find a set of n vertices V_0 of G whose deletion from G results in a disconnected graph. The n vertices of such a set are called *join vertices* of G. The *join graph* $G_0=(V_0, E_0)$ is the induced graph on join vertex set V_0, where E_0 consists of the edges of G having both of their ends in V_0. A join graph is unique for a particular choice of V_0, but there are in general several different sets of join vertices. Let $G_i'=(V_i', E_i')$, $1 \leqq i \leqq k$, be the components of the disconnected graph obtained by deleting the vertices V_0 from G; these components are referred to as the *join components*. Each member of the edge set $(E-E_0-\bigcup_{i=1}^{k} E_i')$ is incident to a vertex of G_0 and a vertex of exactly one G_i', since $V_1', V_2', \cdots, V_k'$ are mutually disjoint. The set $(E-E_0)$ may be partitioned into k subsets $E_1, E_2, \cdots, E_k$ so that each E_i is the union of E_i' and the edges incident to the vertices of G_i'. Then $G_i=G \cdot E_i$, $1 \leqq i \leqq k$, are connected and called the *n-palms* of G associated with the join vertex set V_0.

LEMMA 1. *If the vertex connectivity of a graph* $G=(V, E)$ *is* $n(\geqq 1)$, *then* $\lambda(G) \leqq n$.

Proof. If n is infinity, the lemma is trivial. Suppose n is finite. We show that there exists a nonnull proper subset S of E such that $\eta(G; S, \bar{S})=n$ and $\min(r(G \cdot S), r(G \cdot \bar{S})) \geqq n$. Let V_0 be a join vertex set of G. Since the connectivity of G is finite, there are at least two join components associated with V_0, where $|V_0|=n$. Let G_1' and G_2' be distinct join components of G. Since each of these join components contains at least one vertex, the corresponding n-palms $G_1=(V_1, E_1)$ and $G_2=(V_2, E_2)$ both contain at least $n+1$ vertices of G. Choose $S=E_1$ and $\bar{S}=E-E_1 \supseteq E_2$. Clearly, $r(G \cdot S)=r(G \cdot E_1) \geqq n$ and $r(G \cdot \bar{S}) \geqq r(G \cdot E_2) \geqq n$. In addition, the members of V_0 are the only common vertices of $G \cdot E_1$ and $G \cdot (E-E_1)$. Thus $\eta(G; S, \bar{S})=|V_0|=n$, and the lemma follows. □

The other half of Theorem 1 is stated as follows.

LEMMA 2. *If the vertex connectivity of a graph* $G=(V, E)$ *is* $n(\geqq 1)$, *then* $\lambda(G) \geqq n$.

Proof. If $\lambda(G)=\infty$, it is obvious that the lemma is true. To prove the lemma for a finite λ we assume $\lambda(G)=k<n$. By definition, there exists a nonnull proper subset S of E such that $\eta(G; S, \bar{S})=k$ and $\min(r(G \cdot S), r(G \cdot \bar{S})) \geqq k$. First we shall show that $G \cdot S$ and $G \cdot \bar{S}$ are both connected.

Suppose $G \cdot S$ is not connected. Since $r(G \cdot S) \geqq k$, $G \cdot S$ contains at least $k+1$ vertices, and at least one of the vertices, say v, is not a vertex of $G \cdot \bar{S}$, because the number of common vertices of $G \cdot S$ and $G \cdot \bar{S}$ is k. Let $G_1=(V_1, E_1)=G \cdot E_1$ be the component of $G \cdot S$ which contains v. Since $G \cdot (S-E_1)$ has no vertices in common with G_1 and has at least one common vertex with $G \cdot \bar{S}$, we have $\eta(G; E_1, \bar{E}_1)=h<\eta(G; S, \bar{S})=k$, where $\bar{E}_1=E-E_1$. Furthermore, $r(G \cdot E_1)=|V_1|-1 \geqq h$ and $r(G \cdot \bar{E}_1) \geqq r(G \cdot \bar{S}) \geqq k>h$. Hence, $\lambda(G)=h<k$, which is contrary to the hypothesis. Therefore, $G \cdot S$ is connected. Similarly, $G \cdot \bar{S}$ is also connected.

Both $G \cdot S$ and $G \cdot \bar{S}$ contain at least $k+1$ vertices. Since the number of common vertices of $G \cdot S$ and $G \cdot \bar{S}$ is k, the deletion of the common vertices results in a disconnected graph. However, this is impossible since G is n-connected. Therefore, $\lambda(G)=k \geqq n$. □

Theorem 1 is obtained from Lemmas 1 and 2. Originally, the vertex connectivity of a graph was defined as a vertex-removal operation and possessed a strong graph-theoretic flavor. However, Theorem 1 enables one to express this connectivity concept in a more abstract form which can be extended to matroids. In the subsequent sections vertex connectivity will be referred to as the *Whitney connectivity* of graphs, since matroids as a generalization of graphs do not have a vertex concept.

4. Whitney connectivity of matroids. Let $\mathbf{M}=(E, \mathbf{C})$ be a matroid, and let S and $\bar{S}$ be nonnull complementary subsets of E. Define the following function:

$$\xi(\mathbf{M}; S, \bar{S}) = -r(\mathbf{M})+r(\mathbf{M}\times S)+r(\mathbf{M}\times \bar{S})+1.$$

The *Whitney connectivity* (or simply *W-connectivity*) of the matroid $\mathbf{M}$, denoted by $\lambda(\mathbf{M})$, is the least integer n for which there exists an $S \subset E$ such that

$$\xi(\mathbf{M}; S, \bar{S}) = n \quad \text{and} \quad \min\,(r(\mathbf{M}\times S), r(\mathbf{M}\times \bar{S})) \geqq n.$$

If there is no such integer, we then denote $\lambda(\mathbf{M})=\infty$. (Note that the connectivity definition of Tutte replaces the second term by $\min(|S|, |\bar{S}|) \geqq n$.)

The function ξ may be written equivalently in several ways:

$$\begin{aligned}\xi(\mathbf{M}; S, \bar{S}) &= r(\mathbf{M}\times S)-r(\mathbf{M}\cdot S)+1\\ &= -|S|+r(\mathbf{M}\times S)+r(\mathbf{M}^*\times S)+1\\ &= \mu(\mathbf{M})-\mu(\mathbf{M}\times S)-\mu(\mathbf{M}\times \bar{S})+1\\ &= \mu(\mathbf{M}\cdot S)-\mu(\mathbf{M}\times S)+1\\ &= |S|-\mu(\mathbf{M}\times S)-\mu(\mathbf{M}^*\times S)+1.\end{aligned}$$

In the following discussion, use of the same symbol λ for graphs and matroids should not cause any confusion. The next theorem is a consequence of a series of lemmas which follow.

THEOREM 2. *If G is a connected graph, then $\lambda(G)=\lambda(\mathbf{P}(G))$.*

LEMMA 3. *If $G=(V, E)$ is a connected graph and S is a subset of E, then*

$$\xi(\mathbf{P}(G); S, \bar{S}) = \eta(G; S, \bar{S})-c(G\cdot S)-c(G\cdot \bar{S})+2.$$

Proof. By definition,

$$\begin{aligned}\xi(\mathbf{P}(G); S, \bar{S}) &= -r(\mathbf{P}(G))+r(\mathbf{P}(G)\times S)+r(\mathbf{P}(G)\times \bar{S})+1\\ &= -r(\mathbf{P}(G)+r(\mathbf{P}(G\cdot S))+r(\mathbf{P}(G\cdot \bar{S}))+1.\end{aligned}$$

Let V_S and $V_{\bar{S}}$ be the vertex sets of $G\cdot S$ and $G\cdot \bar{S}$, respectively. Since $r(\mathbf{P}(G))=r(G)$, we have

$$\begin{aligned}\xi(\mathbf{P}(G); S, \bar{S}) &= -r(G)+r(G\cdot S)+r(G\cdot \bar{S})+1\\ &= -|V|-c(G)+|V_S|-c(G\cdot S)+|V_{\bar{S}}|-c(G\cdot \bar{S})+1\\ &= -|V|+|V_S|+|V_{\bar{S}}|-c(G\cdot S)-c(G\cdot \bar{S})+2.\end{aligned}$$

Since $|V|=|V_S|+|V_{\bar{S}}|-\eta(G; S, \bar{S})$, we can reduce the above equation to

$$\xi(\mathbf{P}(G); S, \bar{S}) = \eta(G; S, \bar{S})-c(G\cdot S)-c(G\cdot \bar{S})+2. \qquad \square$$

LEMMA 4. *If $G=(V, E)$ is a connected graph, then $\lambda(\mathbf{P}(G)) \leqq \lambda(G)$.*

Proof. Since the lemma is trivial for $\lambda(G)=\infty$, we assume $\lambda(G)=n$, where n is a finite positive integer. For some nonnull proper subset S of E, $\eta(G; S, \bar{S})=n$ and $\min\,(r(G\cdot S), r(G\cdot \bar{S})) \geqq n$. By Lemma 3,

$$\xi(\mathbf{P}(G); S, \bar{S}) = n-c(G\cdot S)-c(G\cdot \bar{S})+2 \leqq n,$$

and $\min\,(r(\mathbf{P}(G)\times S), r(\mathbf{P}(G)\times \bar{S})) \geqq n$. Therefore, $\lambda(\mathbf{P}(G)) \leqq n=\lambda(G)$. $\square$

LEMMA 5. *If $G=(V, E)$ is a connected graph, then $\lambda(\mathbf{P}(G)) \geqq \lambda(G)$.*

Proof. If $\lambda(\mathbf{P}(G))=\infty$, the lemma is obvious.

Suppose $\lambda(\mathbf{P}(G))=n$ is finite and $\lambda(G)>n$. Then, by Lemma 3, there exists a nonnull proper subset S of E such that

$$\eta(G;S,\bar{S})\leqq n+c(G\cdot S)+c(G\cdot\bar{S})-2$$

and $r(G\cdot S)$, $r(G\cdot\bar{S}\geqq n$. We choose S so that the above conditions are satisfied and $\eta(G;S,\bar{S})$ is minimum, consistent with those conditions.

If $c(G\cdot S)+c(G\cdot\bar{S})=2$, then $\lambda(G)\leqq n$, which is contrary to the hypothesis. Therefore, $G\cdot S$ or $G\cdot\bar{S}$ has at least two components. In the following the notation $\eta(S_1,S_2)$ denotes the numbers of common vertices of the two subgraphs $G\cdot S_1$ and $G\cdot S_2$, where S_1 and S_2 are subsets of E. Using this notation we can write $\eta(G;S,\bar{S})=\eta(S,\bar{S})$.

Case 1. One of the components, say $G\cdot S_0$, of $G\cdot S$ or $G\cdot\bar{S}$ satisfies the following condition: $r(G\cdot S_0)\geqq\eta(S_0,\bar{S}_0)$.

Without loss of generality we may suppose that $G\cdot S_0$ is a component of $G\cdot S$. If $\eta(S_0,\bar{S}_0)\leqq n$, then, clearly, $\lambda(G)\leqq\eta(S_0,\bar{S}_0)\leqq n$ since $r(G\cdot\bar{S}_0)\geqq r(G\cdot\bar{S})\geqq n$. This contradicts the hypothesis. If $\eta(S_0,\bar{S}_0)>n$, then let $G\cdot S_0'$ be a component of $G\cdot(S-S_0)$. Let $S'=S-S_0'$ and $\bar{S}'=E-S'=\bar{S}\cup S_0'$. Then

$$\eta(S',\bar{S}')=\eta(S,\bar{S})-\eta(S_0',\bar{S}_0')<\eta(S,\bar{S}),$$

$$c(G\cdot S')=c(G\cdot S)-1,$$

$$c(G\cdot\bar{S}')\geqq c(G\cdot\bar{S})-\eta(S_0',\bar{S}_0')+1.$$

Therefore

$$\begin{aligned}\eta(S',\bar{S}')&\leqq n+c(G\cdot S)+c(G\cdot\bar{S})-2-\eta(S_0',\bar{S}_0')\\&\leqq n+[c(G\cdot S')+1]+[c(G\cdot\bar{S}')+\eta(S_0',\bar{S}_0')-1]-2-\eta(S_0',\bar{S}_0')\\&=n+c(G\cdot S')+c(G\cdot\bar{S}')-2,\end{aligned}$$

$$r(G\cdot\bar{S}')\geqq r(G\cdot\bar{S})\geqq n,$$

$$\begin{aligned}r(G\cdot S')=r(G\cdot S_0)+r(G\cdot(S-(S_0'\cup S_0)))&\geqq r(G\cdot S_0)\\&\geqq\eta(S_0,\bar{S}_0)-1\geqq n.\end{aligned}$$

However, this contradicts the fact that $\eta(S,\bar{S})$ is a minimum.

Case 2. For each component $G\cdot S_0$ of $G\cdot S$ and $G\cdot\bar{S}$, $r(G\cdot S_0)<\eta(S_0,\bar{S}_0)$ and $r(G\cdot S)\geqq n+1$ or $r(G\cdot\bar{S})\geqq n+1$.

Let $G\cdot S_0=(V_0,S_0)$ be any component of $G\cdot S$ or $G\cdot\bar{S}$. Then

$$|V_0|=r(G\cdot S_0)+1\leqq\eta(S_0,\bar{S}_0)\leqq|V_0|.$$

Therefore $\eta(S_0,\bar{S}_0)=|V_0|$, and hence $\eta(S,\bar{S})=|V|$. Suppose $r(G\cdot S)\geqq n+1$, without loss of generality. We shall show that S can be chosen so that $G\cdot S$ contains no polygons. Let S' be a spanning forest of $G\cdot S$. Since the vertex set of $G\cdot S$ is V and $G\cdot S$ contains no isolated vertices, the vertex set of $G\cdot S'$ is V; that is, $\eta(S',\bar{S}')=\eta(S,\bar{S})$. We also have

$$r(G\cdot S')=|S'|=r(G\cdot S)\geqq n+1,$$

$$r(G\cdot\bar{S}')\geqq r(G\cdot S)\geqq n.$$

Thus we can assume that $G\cdot S$ contains no polygons. Let e be an edge of $G\cdot S$ which

has valence one at its one end in $G \cdot S$. Let $S' = S - \{e\}$ and $\bar{S}' = \bar{S} \cup \{e\}$. Then

$$\eta(S', \bar{S}') \leqq \eta(S, \bar{S}) - 1,$$
$$r(G \cdot S') = r(G \cdot S) - 1 \geqq n,$$
$$r(G \cdot \bar{S}') \geqq r(G \cdot \bar{S}) \geqq n.$$

Hence

$$\begin{aligned}\eta(S', \bar{S}') &\leqq n + c(G \cdot S) + c(G \cdot \bar{S}) - 3 \\ &\leqq n + c(G \cdot S') + [c(G \cdot \bar{S}') + 1] - 3 \\ &= n + c(G \cdot S') + c(G \cdot \bar{S}') - 2.\end{aligned}$$

This is contrary to the requirement that $\eta(S, \bar{S})$ be a minimum.

Case 3. For each component $G \cdot S_0$ of $G \cdot S$ and $G \cdot \bar{S}$, $r(G \cdot S_0) < \eta(S_0, \bar{S}_0)$ and $r(G \cdot S) = r(G \cdot \bar{S}) = n$.

By assumption

$$c(G \cdot S) = c(G \cdot \bar{S}) = |V| - n \geqq 2.$$

Suppose the vertex set of a component $G \cdot S_1$ of $G \cdot S$ properly contains the vertex set of a component of $G \cdot \bar{S}$. Let $G \cdot T_i$, $1 \leqq i \leqq k$, be the components of $G \cdot \bar{S}$ whose vertex sets are properly contained in $G \cdot S_1$. Let $S_1' = \bigcup_{i=1}^{k} T_i$, $S_2 = S - S_1$ and $S_2' = \bar{S} - S_1'$. Let $S' = S_1 \cup S_1'$ and $\bar{S}' = S_2 \cup S_2'$. Then

$$r(G \cdot S') \geqq \eta(S', \bar{S}'),$$
$$r(G \cdot \bar{S}') = \eta(S_1, \bar{S}_2') + \eta(S_2, \bar{S}_2') - c(G \cdot \bar{S}').$$

Since every component of $G \cdot \bar{S}'$ contains the vertices of $G \cdot S_2$, we have

$$r(G \cdot \bar{S}') \geqq \eta(S_1, \bar{S}_2') = \eta(S', \bar{S}'),$$
$$\eta(S', \bar{S}') \leqq r(G \cdot S_1) \leqq r(G \cdot S) = n.$$

Accordingly, $\lambda(G) \leqq n$, contrary to the hypothesis.

If $G_1 = (V_1, S_1)$ and $G_2 = (V_2, S_2)$ are components of $G \cdot S$ and $G \cdot \bar{S}$, respectively, and $V_1 = V_2$, then G is not connected, which is a contradiction.

Lastly, we consider the following case: No component of $G \cdot S$ contains the vertices of a component of $G \cdot \bar{S}$, and no component of $G \cdot \bar{S}$ contains the vertices of a component of $G \cdot S$.

Let $G \cdot S_1$ and $G \cdot S_1'$ be components of $G \cdot S$ and $G \cdot \bar{S}$ which have common vertices. Let $S_2 = S - S_1$ and $S_2' = \bar{S} - S_1'$. If we set $S' = S_1 \cup S_1'$ and $\bar{S}' = S_2 \cup S_2'$, then

$$r(G \cdot S') \geqq \eta(S', \bar{S}'),$$
$$r(G \cdot \bar{S}') = \eta(S', \bar{S}') + \eta(S_2, S_2') - c(G \cdot \bar{S}').$$

Every component of $G \cdot \bar{S}'$ contains vertices of $G \cdot S_2$, and every common vertex of $G \cdot S_2$ and $G \cdot S_1'$ is contained in a component of $G \cdot \bar{S}'$, since otherwise the condition of the first part of Case 3 is satisfied. Therefore $\eta(S_2, S_2') \geqq c(G \cdot S_2) \geqq c(G \cdot \bar{S}')$, and $r(G \cdot \bar{S}') \geqq \eta(S', \bar{S}')$. Since $r(G \cdot S) = n = \eta(S', \bar{S}') + \eta(S_1, \bar{S}_1') + \eta(S_2, S_2') - c(G \cdot S_2) - 1$, we have $\eta(S', \bar{S}') \leqq n - \eta(S_2, S_2') + c(G \cdot S_2) \leqq n$. Accordingly, $\lambda(G) \leqq n$, contrary to the hypothesis.

Since all the cases have been examined, we conclude that $\lambda(G) \leqq \lambda(\mathbf{P}(G))$ for a connected graph G. □

Lemmas 4 and 5 establish Theorem 2. □

5. Properties of Whitney connectivity. In the following, comparisons of Whitney connectivity and the connectivity definition of Tutte are presented, and a number of properties of Whitney connectivity are examined.

The concept of matroid connectivity was originally introduced by Tutte [6] as a generalization of another definition of graph connectivity. Let $\mathbf{M} = (E, \mathbf{C})$ be a matroid. The matroid connectivity $\lambda_T(\mathbf{M})$ of $\mathbf{M}$ as defined by Tutte (hereafter called the *T-connectivity* of $\mathbf{M}$) is the least integer n for which there exists a subset $S \subset E$ such that $\xi(\mathbf{M}; S, \bar{S}) = n$ and $\min(|S|, |\bar{S}|) \geqq n$. If there is no such integer, we then write $\lambda_T(\mathbf{M}) = \infty$. The main difference between the two definitions of connectivity are the following constraints:

$$\text{Whitney:} \quad \min(r(\mathbf{M}\times S), r(\mathbf{M}\times \bar{S})) \geqq n,$$

$$\text{Tutte:} \quad \min(|S|, |\bar{S}|) \geqq n.$$

Since $r(\mathbf{M}\times S) \leqq |S|$ and $r(\mathbf{M}\times \bar{S}) \leqq |\bar{S}|$, we have $\lambda_T(\mathbf{M}) \leqq \lambda(\mathbf{M})$ for any matroid $\mathbf{M}$.

Let $\mathbf{M} = (E, \mathbf{C})$ be a matroid. Then $\mathbf{M}$ is *n-connected* if $2 \leqq n \leqq \lambda(\mathbf{M})$ and is *connected* if $\lambda(\mathbf{M}) \geqq 2$. If $\lambda(\mathbf{M}) = 1$, $\mathbf{M}$ is said to be *separable*. A nonnull proper subset S of E is an *n-separator* if $\lambda(\mathbf{M}) = n$, $\xi(\mathbf{M}; S, \bar{S}) = n$ and $\min(r(\mathbf{M}\times S), r(\mathbf{M}\times \bar{S})) \geqq n$. A 1-separator is also called a *separator.*

This terminology may also be defined in Tutte's sense. In the above definitions, substitute $\lambda_T(\mathbf{M})$ for $\lambda(\mathbf{M})$ and $\min(|S|, |\bar{S}|)$ for $\min(r(\mathbf{M}\times S), r(\mathbf{M}\times \bar{S}))$; then the corresponding concepts are respectively called *T-n-connected, T-connected, T-separable, T-n-separator* and *T-separator.*

The following theorem provides a condition for the two definitions of separators to be equivalent.

THEOREM 3. *Let $\mathbf{M} = (E, \mathbf{C})$ be a matroid containing no loops and S be a nonnull proper subset of E. Then S is a separator of $\mathbf{M}$ if and only if it is a T-separator of $\mathbf{M}$.*

Although $\lambda_T(\mathbf{M}) \leqq \lambda(\mathbf{M})$ is true in general, the Tutte connectivity of $\mathbf{M}$ is much smaller than the Whitney connectivity in many interesting cases. Consider the polygon matroid of the complete bipartite graph $K_{n,n}$, where $n \geqq 4$. The Tutte connectivity of this matroid is $\lambda_T(\mathbf{P}(K_{n,n})) = 4$, while $\lambda(\mathbf{P}(K_{n,n})) = n$. For a complete graph K_n, $n \geqq 4$, we have $\lambda_T(\mathbf{P}(K_n)) = 3$ and $\lambda(\mathbf{P}(K_n)) = \infty$. A necessary and sufficient condition for $\lambda(\mathbf{M}) = \lambda_T(\mathbf{M})$ is given in the next theorem.

THEOREM 4. *Let $\mathbf{M} = (E, \mathbf{C})$ be a matroid of T-connectivity n, where n is finite; then $\lambda(\mathbf{M}) = \lambda_T(\mathbf{M})$ if and only if $\mathbf{M}$ has a T-n-separator S such that neither S nor $\bar{S}$ contains a base of $\mathbf{M}$.*

Proof. Suppose $\lambda(\mathbf{M}) = \lambda_T(\mathbf{M}) = n$. Let S be an n-separator of $\mathbf{M}$. Then

$$\xi(\mathbf{M}; S, \bar{S}) = n,$$

$$\min(r(\mathbf{M}\times S), r(\mathbf{M}\times \bar{S})) \geqq n.$$

Since $r(\mathbf{M}\times S) \leqq |S|$ and $r(\mathbf{M}\times \bar{S}) \leqq |\bar{S}|$, S is a T-n-separator of $\mathbf{M}$. By assumption,

$$-r(\mathbf{M}) + r(\mathbf{M}\times S) + r(\mathbf{M}\times \bar{S}) + 1 \leqq r(\mathbf{M}\times S), r(\mathbf{M}\times \bar{S}),$$

or,

$$r(\mathbf{M}\times S), r(\mathbf{M}\times \bar{S}) \leqq r(\mathbf{M}) - 1.$$

Therefore neither S nor $\bar{S}$ contains a base of $\mathbf{M}$.

Now suppose S is a T-n-separator of $\mathbf{M}$ such that neither S nor $\bar{S}$ contains a base of $\mathbf{M}$. Then

$$r(\mathbf{M}\times S),\ r(\mathbf{M}\times\bar{S})\leqq r(\mathbf{M})-1,$$

and

$$\begin{aligned}\xi(\mathbf{M};S,\bar{S})=n&=-r(\mathbf{M})+r(\mathbf{M}\times S)+r(\mathbf{M}\times\bar{S})+1\\&\leqq r(\mathbf{M}\times S),\ r(\mathbf{M}\times\bar{S}).\end{aligned}$$

Thus $\lambda(\mathbf{M})\leqq n$, and hence $\lambda(\mathbf{M})=\lambda_T(\mathbf{M})=n$ since $\lambda_T(\mathbf{M})\leqq\lambda(\mathbf{M})$. □

A number of additional properties of Whitney connectivity are stated in the subsequent theorems. In [2] the class of matroids with nonfinite T-connectivity, i.e., $\lambda_T(\mathbf{M})=\infty$, is identified as being a subclass of binomial or k-uniform matroids. The next theorem characterizes such matroids for Whitney connectivity.

THEOREM 5. *The following statements are equivalent*:

(a) $\lambda(\mathbf{M})=\infty$.

(b) *For each nonnull proper subset S of E, S or $\bar{S}$ contains a base of* $\mathbf{M}$.

(c) $r(\mathbf{M}\cdot S)=0$ *or* $r(\mathbf{M}\cdot\bar{S})=0$ *for each nonnull proper subset S of E.*

(d) *For each cocircuit C^* of* $\mathbf{M}$, $E-C^*$ *contains no cocircuits of* $\mathbf{M}$.

Proof. (a)⇔(b). By assumption, for each nonnull proper subset S of E,

$$\begin{aligned}\xi(\mathbf{M};S,\bar{S})&=-r(\mathbf{M})+r(\mathbf{M}\times S)+r(\mathbf{M}\times\bar{S})+1\\&\geqq r(\mathbf{M}\times S)+1\text{ or }r(\mathbf{M}\times\bar{S})+1.\end{aligned}$$

The following condition is equivalent to the one above:

$$r(\mathbf{M}\times\bar{S})\geqq r(\mathbf{M})\quad\text{or}\quad r(\mathbf{M}\times S)\geqq r(\mathbf{M}).$$

Since $r(\mathbf{M}\times S),\ r(\mathbf{M}\times\bar{S})\leqq r(\mathbf{M})$, we have

$$r(\mathbf{M}\times\bar{S})=r(\mathbf{M})\quad\text{or}\quad r(\mathbf{M}\times S)=r(\mathbf{M}).\tag{1}$$

Accordingly, S or $\bar{S}$ contains a base of $\mathbf{M}$.

If (b) is true, that is, if condition (1) holds, then for each nonnull proper subset S of E

$$\xi(\mathbf{M};S,\bar{S})=r(\mathbf{M}\times S)+1\text{ or }r(\mathbf{M}\times\bar{S})+1,$$

and (a) follows.

(b)⇔(c). If $r(\mathbf{M}\times T)=r(\mathbf{M})$, where $T\subset E$, then $r(\mathbf{M}\cdot\bar{T})=r(\mathbf{M})-r(\mathbf{M}\times T)=0$. Therefore condition (1) is equivalent to condition (2):

$$r(\mathbf{M}\cdot S)=0\quad\text{or}\quad r(\mathbf{M}\cdot\bar{S})=0.\tag{2}$$

(c)⇔(d). Since $\mu(\mathbf{M}^*\times T)=r(\mathbf{M}\cdot T)$, for $T\subset E$ we have the following equivalent condition to (2):

$$\mu(\mathbf{M}^*\times S)=0\quad\text{or}\quad\mu(\mathbf{M}^*\times\bar{S})=0.$$

If C^* is a cocircuit of $\mathbf{M}$, then $\mu(\mathbf{M}^*\times C^*)=1$, and hence $\mu(\mathbf{M}^*\times\bar{C}^*)=0$. Thus $\bar{C}^*=E-C^*$ contains no circuits of $\mathbf{M}^*$, or equivalently, no cocircuits of $\mathbf{M}$.

Suppose (d) is satisfied. If a nonnull proper subset S or E contains a cocircuit of $\mathbf{M}$, then $\mu(\mathbf{M}^*\times\bar{S})=0$ by assumption. If S does not contain cocircuits of $\mathbf{M}$, then $\mu(\mathbf{M}^*\times S)=0$. Thus, for each nonnull proper subset S of E, we have

$$\mu(\mathbf{M}^*\times\bar{S})=0\quad\text{or}\quad\mu(\mathbf{M}^*\times S)=0.$$

Accordingly, (c) follows. □

From Theorem 5 we can identify all the graphs with infinite connectivity.

COROLLARY 1. *Let G be a connected graph containing neither loops nor parallel edges. Then $\lambda(G)=\infty$ if and only if G is a complete graph.*

Proof. If G is a complete graph, clearly $\lambda(G)=\infty$. Suppose $\lambda(G)=\infty$, By Theorem 2, $\lambda(\mathbf{P}(G))=\lambda(G)=\infty$. The dual matroid of $\mathbf{P}(G)$ consists of all the cut-sets of G. Let v and v' be any distinct vertices of G, and S and S' be the star cut-sets at v and v', respectively. These star cut-sets are uniquely determined for the given vertices since G is not separable. According to Theorem 5(d), $E-S$ contains no star cut-sets of G; hence, $S\cap S'\neq\varnothing$. Consequently, v and v' are adjacent. Since G contains no parallel edges, $S\cap S'$ consists of a single element. Therefore any two distinct vertices of G are connected by one edge and consequently G is a complete graph, for G contains no loops. □

Since graph connectivity is a special case of matroid connectivity, the next theorem is obvious.

THEOREM 6. *For a given finite positive integer n, there exists a matroid $\mathbf{M}$ such that $\lambda(\mathbf{M})=n$.*

The following theorem follows from Theorem 3.

THEOREM 7. *Let $\mathbf{M}=(E,\mathbf{C})$ be a matroid containing neither loops nor isthmuses. Then $\mathbf{M}$ is connected if and only if $\mathbf{M}^*$ is connected.*

In T-connectivity the connectivity of a matroid is the same as that of the dual matroid; i.e., $\lambda_T(\mathbf{M})=\lambda_T(\mathbf{M}^*)$. However, this is not true for W-connectivity. The W-connectivity of the polygon matroid of graph G, shown in Fig. 1, is two; however, the W-connectivity of the bond matroid of G, which is the polygon matroid of G^* shown in Fig. 2, is arbitrarily high. Two sufficient conditions for $\lambda(\mathbf{M})\geqq\lambda(\mathbf{M}^*)$ may be stated in terms of circuits and cocircuits.

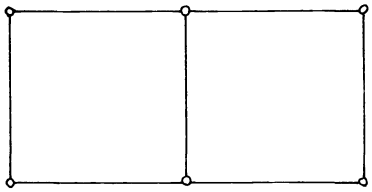


FIG. 1. *Graph G.*

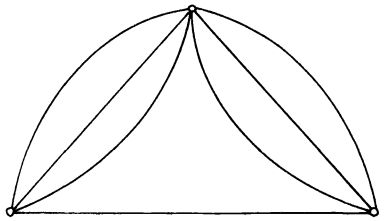


FIG 2. *Graph G*.*

In graph theory it is a well-known fact that for a graph G with connectivity n, the number of edges of every cut-set of G is at least n, and in particular the valence of every vertex of G is at least n. The following lemma is a matroid generalization of this property.

LEMMA 6. *Let $\mathbf{M}=(E,\mathbf{C})$ be a matroid and $\mathbf{M}^*=(E,\mathbf{C}^*)$ be its dual. If $\lambda(\mathbf{M})=n$ is finite, then $r(\mathbf{M})\geqq n+1$, and $r(\mathbf{M}\times C^*)\geqq n$ for each $C^*\in\mathbf{C}^*$.*

Proof. By hypothesis, there exists a nonnull proper subset S of E such that

$$\xi(\mathbf{M}; S, \bar{S}) = -r(\mathbf{M}) + r(\mathbf{M}\times S) + r(\mathbf{M}\times \bar{S}) + 1 = n,$$

$$\min\,(r(\mathbf{M}\times S), r(\mathbf{M}\times \bar{S})) \geqq n.$$

Then

$$r(\mathbf{M}) = r(\mathbf{M}\times S) + r(\mathbf{M}\times \bar{S}) + 1 - n$$

$$\geqq n + n + 1 - n = n + 1.$$

The second part of the lemma is obtained by using the first part. Suppose there exists a cocircuit C^* of $\mathbf{M}$ such that $r(\mathbf{M}\times C^*) \leqq n-1$. Since $r(\mathbf{M}\times \bar{C}^*) = r(\mathbf{M}) - r(\mathbf{M}\cdot C^*) = r(\mathbf{M}) - \mu(\mathbf{M}^*\times C^*) = r(\mathbf{M}) - 1$, we have

$$\xi(\mathbf{M}; C^*, \bar{C}^*) = r(\mathbf{M}\times C^*) \leqq n-1.$$

From the first part of this lemma

$$r(\mathbf{M}\times \bar{C}^*) = r(\mathbf{M}) - \mu(\mathbf{M}^*\times C^*) \geqq n + 1 - 1 = n.$$

Therefore $\lambda(\mathbf{M}) \leqq n-1$, contrary to the hypothesis. Accordingly, $r(\mathbf{M}\times C^*) \geqq n$ for each $C^* \in \mathbf{C}^*$. □

THEOREM 8. *Let* $\mathbf{M} = (E, \mathbf{C})$ *be a matroid of W-connectivity n, where n is finite, and* $\mathbf{M}^* = (E, \mathbf{C}^*)$ *be its dual. The following are then true*:

(a) *If* $\min_{C^*\in\mathbf{C}^*} |C^*| \geqq n+1$, *then* $\lambda(\mathbf{M}^*) \leqq n$.

(b) *If* $\min_{C\in\mathbf{C}} |C| \leqq n-1$ *and* $\mu(\mathbf{M}) \geqq n$, *then* $\lambda(\mathbf{M}^*) \leqq n-1$.

Proof. (a) Since $\lambda(\mathbf{M}) = n$, there exists a nonnull proper subset S of E such that

$$\xi(\mathbf{M}; S, \bar{S}) = n,$$

$$\min\,(r(\mathbf{M}\times S), r(\mathbf{M}\times \bar{S})) \geqq n.$$

If S contains a circuit of M^*, then

$$r(\mathbf{M}^*\times S) \geqq \min_{C^*\in\mathbf{C}^*} |C^*| - 1 \geqq n.$$

Suppose S does not contain circuits of $\mathbf{M}^*$. Then

$$r(\mathbf{M}^*\times S) = |S| \geqq r(\mathbf{M}\times S) \geqq n.$$

In both cases we have $r(\mathbf{M}^*\times S) \geqq n$. Similarly, $r(\mathbf{M}^*\times \bar{S}) \geqq n$ for every nonnull subset S of E. Accordingly, $\lambda(\mathbf{M}^*) \leqq n$.

(b) Let C be a circuit of $\mathbf{M}$ satisfying $|C| \leqq n-1$. Then

$$\xi(\mathbf{M}; C, \bar{C}) = |C| - \mu(\mathbf{M}\times C) - \mu(\mathbf{M}^*\times C) + 1$$

$$= |C| - \mu(\mathbf{M}^*\times C).$$

By Lemma 6 $|C^*| \geqq n$ for every member of $\mathbf{C}^*$, and C does not contain circuits of $\mathbf{M}^*$. Hence

$$\xi(\mathbf{M}; C, \bar{C}) = |C| \quad \text{and} \quad r(\mathbf{M}^*\times C) = |C|.$$

We also have

$$r(\mathbf{M}^*\times \bar{C}) = \mu(\mathbf{M}\cdot \bar{C}) = \mu(\mathbf{M}) - \mu(\mathbf{M}\times C)$$

$$= \mu(\mathbf{M}) - 1 \geqq n-1$$

$$\geqq |C|.$$

Therefore $\lambda(\mathbf{M}^*) \leqq |C| \leqq n-1$. □

A special case of this theorem is the next graph theorem.

COROLLARY 2. *Let $G=(V, E)$ be a planar connected graph of W-connectivity n, where n is finite, and G^* be its dual. We then have*:

(a) *If the minimum cardinality of the cut-sets of G is greater than n, the connectivity of G^* is at most n.*

(b) *If the minimum cardinality of the polygons of G is less than n and $|E|-|V|+1 \geqq n$, then the connectivity of G^* is at most $n-1$.*

In the following theorem we will give a simple sufficient condition for $\lambda(\mathbf{M}) = \lambda(\mathbf{M}^*)$.

THEOREM 9. *Let $\mathbf{M}=(E, \mathbf{C})$ be a matroid and $\mathbf{M}^*=(E, \mathbf{C}^*)$ be its dual. If*

$$\min_{C \in \mathbf{C}} |C| = \min_{C^* \in \mathbf{C}^*} |C^*| = n,$$

and $r(\mathbf{M}), \mu(\mathbf{M}) \geqq n+1$, then $\lambda(\mathbf{M}) = \lambda(\mathbf{M}^*) \leqq n$.

Proof. Step 1. We prove $\lambda(\mathbf{M}), \lambda(\mathbf{M}^*) \leqq n$. Suppose C^* is a circuit of $\mathbf{M}^*$ such that $|C^*| = n$. Then

$$\begin{aligned}\xi(\mathbf{M}; C^*, \bar{C}^*) &= |C^*| - \mu(\mathbf{M} \times C^*) - \mu(\mathbf{M}^* \times C^*) + 1 \\ &= r(\mathbf{M} \times C^*) \leqq |C^*| = n.\end{aligned}$$

However,

$$\begin{aligned}r(\mathbf{M} \times \bar{C}^*) &= r(\mathbf{M}) - r(\mathbf{M} \cdot C^*) \\ &= r(\mathbf{M}) - \mu(\mathbf{M}^* \times C^*) = r(\mathbf{M}) - 1 \geqq n.\end{aligned}$$

Thus $\lambda(\mathbf{M}) \leqq n$. Similarly, we can show $\lambda(\mathbf{M}^*) \leqq n$.

Step 2. We prove $\lambda(\mathbf{M}) = \lambda(\mathbf{M}^*)$.

We assume that $\lambda(\mathbf{M}) = k$ and $\lambda(\mathbf{M}^*) = k^*$, where $k, k^* \leqq n$, as shown above. Let $\mathbf{S}$ and $\mathbf{S}^*$ be the collections of k-separators of $\mathbf{M}$ and k^*-separators of $\mathbf{M}^*$, respectively:

$$\mathbf{S} = \{(S, \bar{S}) | \xi(\mathbf{M}; S, \bar{S}) = k \text{ and } r(\mathbf{M} \times S), r(\mathbf{M} \times \bar{S}) \geqq k\}.$$

$$\mathbf{S}^* = \{(S^*, \bar{S}^*) | \xi(\mathbf{M}^*; S^*, \bar{S}^*) = k^* \text{ and } r(\mathbf{M}^* \times S), r(\mathbf{M}^* \times \bar{S}^*) \geqq k^*\}.$$

If $\mathbf{S} \cap \mathbf{S}^*$ is not null, then there exists $(S, \bar{S}) \in \mathbf{S} \cap \mathbf{S}^*$ and

$$k = \xi(\mathbf{M}; S, \bar{S}) = \xi(\mathbf{M}^*; S, \bar{S}) = k^*.$$

Accordingly, $\lambda(\mathbf{M}) = \lambda(\mathbf{M}^*)$.

Suppose $\mathbf{S} \cap \mathbf{S}^* = \varnothing$. Then, if $(S, \bar{S})$ is a member of $\mathbf{S}$, $r(\mathbf{M}^* \times S) \leqq k-1$ or $r(\mathbf{M}^* \times \bar{S}) \leqq k-1$. Without loss of generality, we assume $r(\mathbf{M}^* \times S) \leqq k-1$. Then we have the following equality:

$$\begin{aligned}\mu(\mathbf{M} \times S) &= -\xi(\mathbf{M}; S, \bar{S}) + \mu(\mathbf{M} \cdot S) + 1 \\ &= -k + r(\mathbf{M}^* \times S) + 1 \leqq -k + (k-1) + 1 \\ &= 0,\end{aligned}$$

and similarly

$$\begin{aligned}\mu(\mathbf{M}^* \times S) &= -\xi(\mathbf{M}^*; S, \bar{S}) + \mu(\mathbf{M}^* \cdot S) + 1 \\ &= -k + r(\mathbf{M} \times S) + 1 \\ &= 1.\end{aligned}$$

Therefore S contains no circuits of $\mathbf{M}$; however, it contains a circuit of $\mathbf{M}^*$. Let C^* be a

circuit of $\mathbf{M}^*$ contained in S. Since $r(\mathbf{M}^* \times S) \geqq r(\mathbf{M}^* \times C^*)$, we have

$$\xi(\mathbf{M}; S, \bar{S}) = k = \mu(\mathbf{M} \cdot S) - \mu(\mathbf{M} \times S) + 1$$

$$= \mu(\mathbf{M} \cdot S) + 1 = r(\mathbf{M}^* \times S) + 1$$

$$\geqq r(\mathbf{M}^* \times C^*) + 1 = |C^*| \geqq \min_{C^* \in \mathbf{C}^*} |C^*| = n.$$

Consequently, $\lambda(\mathbf{M}) = k = n \geqq \lambda(\mathbf{M}^*)$.

Since $\mathbf{S}^*$ is not null, we now choose $(S^*, \bar{S}^*) \in \mathbf{S}^*$. Repeating a similar discussion to the above, we obtain $\lambda(\mathbf{M}^*) = k^* = n$. Therefore $\lambda(\mathbf{M}) = \lambda(\mathbf{M}^*) = n$, and the proof is complete. □

In the next corollary we state the corresponding graph theorem.

COROLLARY 3. *Let $G = (V, E)$ be a planar connected graph and let $|V| \geqq n + 2$ and $|E| \geqq |V| + n$. If the minimum cardinality of the polygons of G is n, which is also the minimum cardinality of the cut-sets, then the connectivity of G is equal to that of a dual graph G^*.*

Proof. Since G is connected, we have

$$r(G) = |V| - 1 \geqq n + 1,$$

$$\mu(G) = |E| - r(G) \geqq n + 1.$$

The corollary follows from Theorem 9. □

A number of properties of Whitney connectivity of matroids have been explored in this section. The authors strongly believe that other theorems on vertex connectivity may be generalized to matroids. Previously, the connectivity definition of Tutte was used in generalizing graph theorems to matroids: for instance, Menger's theorem for matroids [5], matroid decomposition and graph-realizability of matroids [1], [3]. We may also attempt to state these and other theorems in terms of Whitney connectivity. In addition, those theorems on Tutte connectivity which are based only on the connectivity function will also apply to Whitney connectivity. For example, since $\lambda(M) \leqq \max \xi(\mathbf{M}; S, \bar{S})$ for a matroid of finite connectivity, a trivial upper bound on the connectivity is given by $\min(r(\mathbf{M}) + 1, \mu(\mathbf{M}) + 1)$. We have also found the maximum value of ξ in terms of the distance between maximally distant bases, a concept useful for many other applications. If B_1 and B_2 are bases of $\mathbf{M}$, then the distance between them is defined by $|B_1 - B_2|$. A pair of bases which have the maximum distance are called *maximally distant bases.* If we let B_1 and B_2 be a pair of maximally distant bases and define $d_0 = |B_1 - B_2|$, where d_0 may be easily computed, then we have shown [2]

$$\max_{S \subseteq E} \xi(\mathbf{M}; S, \bar{S}) = d_0 + 1.$$

Thus the connectivity is never greater than $d_0 + 1$.

We conclude with the statement of an unsolved problem related to the connectivity function ξ. Let B be a base of a matroid $\mathbf{M}$ and $\bar{B}$ its complement. The following question arises in applications: what is the minimum value of $r(\mathbf{M} \times \bar{B})$ for all the bases B of $\mathbf{M}$? Since $\xi(\mathbf{M}; B, \bar{B}) = r(\mathbf{M} \times \bar{B}) + 1$, this problem is equivalent to finding the value of $\min_{S:\text{base}} \xi(\mathbf{M}; S, \bar{S})$. The corresponding graph theory problem is called the *central tree* problem: given any graph G, what is the minimum rank of spanning roses (cotrees)

of G? This problem is unsolved, and at present there exists no efficient algorithm for finding a spanning tree that satisfies the above condition.

REFERENCES

[1] R. E. BIXBY, *Composition and decomposition of matroids and related topics*, Tech. Rep. 147, Department of Operations Research, Cornell University, Ithaca, NY, 1972.
[2] T. INUKAI AND L. WEINBERG, *Theorems on matroid connectivity*, Discrete Math., 22 (1978), pp. 311–312.
[3] ———, *Graph realizability of matroids*, Ann. New York Acad. Sci., 319 (1979), pp. 289–305.
[4] W. T. TUTTE, *A theory of 3-connected graphs*, Proc. 64 Nederl. Akad. Wetensch., 1961, pp. 441–455.
[5] ———, *Menger's theorem for matroids*, J. Res. NBS, 69B(1965), pp. 49–53.
[6] ———, *Connectivity in matroids*, Canad. J. Math., 18 (1966), pp. 1301–1324.
[7] D. J. A. WELSH, *Matroid Theory*, Academic Press, New York, 1976.
[8] H. WHITNEY, *Congruent graphs and the connectivity of graphs*, Amer. J. Math., 54(1932), pp. 150–168.

SIAM. J. ALG. DISC. METH.
Vol. 2, No. 2, June 1981

0196-5212/81/0202-0005 $01.00/0

SOME CONSTRUCTIONS FOR CONVOLUTIONAL AND BLOCK CODES*

PIERRE A. VON KAENEL†

Abstract. A new construction for a class of linear block codes and an extension of the construction to a class of convolutional codes are presented. Lower and upper bounds on the minimum distance and free distance for the constructed codes are determined. In the case of the convolutional codes, a criterion for constructing a noncatastrophic encoder is given.

1. Introduction. The main result of this paper is presented in § 3 and consists of the construction of noncatastrophic encoders of the form $N(D)=N_1+N_2D$ which generate a new class of convolutional codes of block length two. Lower and upper bounds on the free distance of these codes are also determined. The block codes defined as the row space of $[N_1N_2]$ have been analyzed, and, because a number of good binary block codes have been found, a general construction for these codes is presented in § 2.

2. A class of linear block codes. Let G_i be an $a_i \times n$ matrix over GF(2) having rank a_i $(i=1,2)$, where $a_2>1$, and let C_i be its row space. Choose T, an $a_2 \times a_2$ nonsingular matrix over GF(2) with the property that $uT \neq u$ holds for all nonzero binary a_2-tuples u. For example, T may be chosen as the companion matrix of a primitive polynomial of degree a_2 over GF(2).

Construction K_1. Let C denote the $(2n, a_1+a_2)$ code defined as the row space of M, where

$$M=[M_1, M_2],$$

and

$$M_1=\begin{bmatrix} G_1 \\ G_2 \end{bmatrix}, \qquad M_2=\begin{bmatrix} G_1 \\ TG_2 \end{bmatrix}.$$

(That M has dimension a_1+a_2 is a consequence of Theorem 1.)

If C_1 and C_2 are cyclic codes, and C_2 is not generated by the all one vector, then C is a quasicyclic code.

Bounds on $d(C)$, the minimum distance of a code derived from construction K_1, are next determined as a function of the minimum distances $d(C_1)$ and $d(C2)$.

THEOREM 1. *The minimum distance* $d(C)$ *satisfies*

$$\min\{2d(C_1), d(C_2)\} \leqq d(C) \leqq 2d(C_1).$$

Proof. The minimum distance of C is determined by considering the weight of codeword xM, where $x=(x_1, x_2)$ and x_i is a row a_i-tuple.

Case 1. $x_1 \neq 0$, $x_2=0$. Then $w(xM) \geqq 2d(C_1)$, where equality holds for the proper choice of x_1.

Case 2. $x_2 \neq 0$. Then

$$\begin{aligned} w(xM) &= w(xM_1)+w(xM_2) \\ &\geqq w(x(M_1+M_2)) \\ &= w(x_2(G_2+TG_2)) \\ &\geqq d(C_2). \end{aligned}$$

□

* Received by the editors July 25, 1978 and in revised form October 2, 1980.

† Department of Mathematics and Computer Science, University of Nebraska at Omaha, Omaha, Nebraska 68101.

The parameters n and k of binary linear codes that have good minimum distance are listed in the Appendix.

3. A class of convolutional codes.

3.1. Definitions and construction. The following construction is an extension of K_1 to convolutional codes of length two. Unfortunately, designing these codes with good distance properties is more complex than designing good block codes. As a result the block codes C_1 and C_2 used to generate the convolutional codes cannot be chosen arbitrarily as in construction K_1.

Let P be an $n \times n$ nonsingular matrix, and choose M_1, an $(a_1+a_2)\times n$ matrix of rank (a_1+a_2), satisfying the following conditions:

1. If B and BP denote the row spaces of M_1 and M_1P, respectively, then $B \neq BP$.
2.

$$M_1 = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix},$$

where G_i is an $a_i \times n$ matrix of rank a_i $(i=1,2)$ and G_1 generates $B \cap BP$. If $B \cap BP = \{0\}$, G_1 is chosen to generate any proper subspace of B.

Denote by C_1, C_2 and C_1P^h the row spaces of G_1, G_2 and G_1P^h (h an integer), respectively. Next define T, an $a_2 \times a_2$ nonsingular matrix over GF(2) satisfying

(i) $$uT \neq u,$$

and, if $B \cap BP \neq \{0\}$,

(ii) $$uTG_2 \in C_1P^{-1} \Rightarrow uG_2 \in C_1P^{-1},$$

for all nonzero binary a_2-tuples u. Finally we define matrix M_2 so that

$$M_2 = \begin{bmatrix} G_1 \\ TG_2 \end{bmatrix}.$$

Construction K_2. Let C_D denote the convolutional code generated by $M(D) = M_1 + (M_2P)D$.

Implementing K_2 may appear difficult, since the matrices P, G_1, G_2 and T are interrelated, especially when $B \cap BP \neq \{0\}$. Furthermore, the resulting encoder, $M(D)$, may even be catastrophic. This occurs when an infinite information sequence yields a finite codeword. We address these difficulties and then determine bounds on the free distance of C_D in the next two subsections.

3.2. Noncatastrophic encoders. The following theorem gives a criterion for a noncatastrophic encoder.

THEOREM 2. *Let C_D be the convolutional codes defined above. $M(D)$ is a noncatastrophic encoder if and only if*

$$\bigcap_{h=0}^{N} C_1P^h = \{0\} \quad \textit{for some } N < \infty.$$

Proof. The theorem is a consequence of the discussion found in [2, § 3]. We need only consider the case $B \cap BP \neq \{0\}$. If $M(D)$ is catastrophic, then there exists an infinite sequence of nonzero (a_1+a_2)-tuples u_i that satisfy $u_iM_2P + u_{i+1}M_1 = 0$ for all $i \geqq 0$. Since $B \cap BP = C_1$, then

(1) $$u_iM_1,\ u_iM_2P \in C_1, \qquad i > 0.$$

Assume that

$$u_i M_1 \in \bigcap_{h=0}^{i-1} C_1 P^h, \qquad i>0 \tag{2}$$

holds. Then $u_i M_1 = u_i M_2$, and from (1) we have

$$u_i M_2 P \in \bigcap_{h=0}^{i} C_1 P^h,$$

which implies

$$u_{i+1} M_1 \in \bigcap_{h=0}^{i} C_1 P^h$$

for the nonzero n-tuple $u_{i+1}M_1$. However, (2) holds for $i=1$; hence by induction we conclude that, for any $N<\infty$,

$$\bigcap_{h=0}^{N} C_1 P^h \neq \{0\}.$$

Conversely, if $\bigcap_{h=0}^{\infty} C_1 P^h = E \neq \{0\}$, then $u_i M_1 \in E \Rightarrow u_i M_2 \in E \Rightarrow u_i M_2 P \in E$ for nonzero n-tuples $u_i M_1$, $u_i M_2$ and $u_i M_2 P$. Given $u_i M_2 P \in E$, there exists u_{i+1} such that $u_i M_2 P + u_{i+1} M_1 = 0$. This construction implies $M(D)$ is catastrophic. □

3.3. Cyclic convolutional codes. If B and BP are equivalent cyclic codes which do not contain the all one vector, then C_D is a generalization of a cyclic convolutional code (CCC). The algebraic structure of these codes can be used to choose G_1, G_2, and T without too much difficulty.

A binary CCC [3] is a code generated by sequences in the form

$$g(X, D) = \sum_{j=0}^{m-1} D^j e(X^{\pi^j})[f(x^{\pi^j})]^{b(j)} \bmod (X^n - 1)$$

where $e(x)$ is the idempotent generator of an irreducible (n, k, d) cyclic block code, whose parity-check polynomial is $h(X)$, and where $f(X)$ is a primitive polynomial for the field of polynomials modulo $h(X)$. The convention $b(j)=\infty$ is used when the coefficient of D^j is 0. Also, $(\pi, n)=1$, $(n, 2)=1$, $\pi \neq 1$ and $b(0)\neq\infty$. These convolutional codes appear to have as rich an algebraic structure as cyclic block codes; however, little has been done in devising algebraic constructions for these codes.

Since $e(X)$ and $e(X^{\pi^j})$ generate equivalent cyclic codes, construction K_2 may be applied to generate encoders of length 2 as follows. First choose a set of distinct primitive idempotent polynomials $E=\{e_i(X)|i=1, \cdots, m\}$ for which $\bigcap_{j=0}^{N}\{e_i(X^{\pi^j})|i=1, \cdots, m\}=\varnothing$ for some $N<\infty$. If B is the cyclic code generated by E, and matrix P defines the automorphism $\Phi: f(X)\to f(X^{\pi})$, then C_1 is the code generated by $E \cap EP (EP=\{e_i(X^{\pi})|i=1, \cdots, m\})$ and C_2 is generated by $E-EP$. The resulting encoder $M(D)$ is noncatastrophic by Theorem 2. For the case $B \cap BP \neq \varnothing$, matrix T can be constructed as follows: Let

$$G_2 = \begin{bmatrix} G_2^1 \\ G_2^2 \end{bmatrix},$$

where G_2^1 is a generator matrix of $C_1 P^{-1} \cap C_2$ (the intersection is nonzero, otherwise the encoder is catastrophic). G_2^1 can easily be constructed by using those polynomials $e(x^{\pi^{-1}})$ $(e(x) \in E \cap EP)$ which are contained in $E-EP$. If a_2^i is the rank of G_2^i, $i=1, 2$,

let T_i be an $a_2^i \times a_2^i$ nonsingular matrix with the property that $u_i \neq u_i T_i$ for all nonzero a_2^i-tuples u_i. Then let

$$T = \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix}.$$

3.4. Bounds on $d_f(C_D)$. We now determine bounds on the free distance of C_D given in terms of the minimum distances of the block codes C_1, C_2, B and $B+BP$.

LEMMA 1. *Matrix T defined in construction K_2 satisfies*

$$uM_2 \in C_1 P^{-1} \Rightarrow uM_1 \in C_1 P^{-1}$$

for all nonzero (a_1+a_2)-tuples u.

Proof. Let $u=(u_1, u_2)$ for a_i-tuple u_i. Then if $uM_2 = u_1G_1 + u_2TG_2 \in C_1P^{-1}$ and $uM_1 = u_1G_1 + u_2G_2$, the lemma now follows from condition (ii) of the definition of T. □

THEOREM 3. *If $M(D)$ is a noncatastrophic encoder, and d_f denotes the free distance of C_D, then*

$$\min\{2d(C_1), d(C_2), 2d(B)+d(B+BP)\} \leqq d_f \leqq 2d(C_1).$$

Proof. We consider the weight $w(z)$ of a finite codeword

$$z = u_0M_1 + \sum_{i=1}^{r} [u_{i-1}M_2P + u_iM_1]D^i + u_rM_2PD^{r+1},$$

where u_i is an (a_1+a_2)-tuple.

Case 1.

$$z = u_iM_1D^i + u_iM_2PD^{i+1} \quad \text{for } u_i \neq 0.$$

Since M_2 and M_2P generate equivalent block codes, we have, from Theorem 1, $\min\{2d(C_1), d(C_2)\} \leqq w(z) \leqq 2d(C_1)$.

Case 2.

$$z = u_jM_1D^j + \sum_{i=j+1}^{j+k} [u_{i-1}M_2P + u_iM_1]D^i + u_{j+k}M_2PD^{j+k+1},$$

where $k>0$, u_j, $u_{j+k} \neq 0$ and $u_{i-1}M_2P + u_iM_1 \neq 0$ for some i, $j+1 \leqq i \leqq j+k$. Then $w(z) \geqq d(B)+d(B+BP)+d(B)$.

Case 3.

$$z = u_jM_1D^j + u_{j+k}M_2PD^{j+k+1} \quad \text{for } k>0 \text{ and } u_j, u_{j+k} \neq 0.$$

This case exists only if $B \cap BP \neq \varnothing$. Then $u_{j+k}M_1 \in C_1$ and $u_jM_2P \in C_1$ hold, implying $u_{j+k}M_2P \in C_1P$ and $u_jM_2 \in C_1P^{-1}$. By Lemma 1, $u_jM_1 \in C_1P^{-1}$. Hence $w(z) = w(u_jM_1) + w(u_{j+k}M_2P) \geqq 2d(C_1)$. □

3.5. Examples. In the following examples, $g_i(x)$ represents the irreducible factor of x^n-1 whose roots are $(a^i)^{2^k}$, $k=0, 1, \cdots$, where a is a primitive nth root of unity. The cyclic code generated by a polynomial $f(x)$ is denoted $[f(x)]$,

Example 1. Let $n=31$ and $\pi=3$. Let $B=[g_0(x)g_1(x)g_3(x)]$, a (31, 20, 6) code. Then $BP=[g_0(x^3)g_1(x^3)g_3(x^3)]=[g_0(x)g_{11}(x)g_1(x)]$. $C_1=[g_0(x)g_1(x)g_3(x)g_{11}(x)]$, a (31, 15, 8) code. $C_2=[g_0(x)g_1(x)g_3(x)g_5(x)g_7(x)g_{15}(x)]$, a (35, 5, 16) irreducible code. $B+BP=[g_0(x)g_1(x)]$, a (35, 25, 4) code. If matrix G_1 generates $C_1=B \cap BP$, then

$\cap_{h=0}^{3} C_1P^h = \{0\}$; hence construction K_2 yields a noncatastrophic encoder which generates (by Theorem 3) a (31, 20) CCC with $d_f = 16$.

Example 2. Let $n = 51$ and $\pi = 5$. Let $B = [g_0(x)g_1(x)g_9(x)g_{17}(x)]$, a (51, 32, 6) code. Then $BP = [g_0(x)g_3(x)g_{11}(x)g_{17}(x)]$, $C_1 = [g_0(x)g_1(x)g_3(x)g_9(x)g_{11}(x)g_{17}(x)]$, a (51, 16, 16) code, $C_2 = [g_0(x)g_1(x)g_5(x)g_9(x)g_{17}(x)g_{19}(x)]$, a (51, 16, 14) code, and $B + BP = [g_0(x)g_{17}(x)]$ and has minimum distance at least 2. Since $\cap_{h=0,\cdots,3} C_1P^h = \{0\}$, we have a noncatastrophic encoder generating a (51, 32) CCC with $14 \leqq d_f \leqq 32$.

Appendix. Table 1 lists the parameters of binary linear codes derived by construction K_1 whose lower bounds on the minimum distance d equal and whose upper bounds exceed those of the best known binary linear codes listed in [1]. For (d) a pair of numbers is given. The first number indicates the lower bound on d determined by Theorem 1 which equals the best known d. The second number denotes either the upper bound (from Theorem 1) or the known upper bound on d for any linear (n, k) code (indicated by an asterisk), whichever is smaller. Included in the table is a (98, 31) code with $d = 23$ or 24 which improves the best known (98, 31, 22) code in [1]. There exist many other codes derived from K_1 whose minimum distances equal the best known.

TABLE 1
Linear codes from K_1

n	k	(d)	n	k	(d)	n	k	(d)
24	15	4–5*	80	61	6–8	98	85	4–5*
26	17	4–5*	80	67	4–6	100	32	23–24
28	19	4–5*	82	69	4–6	100	80	6–8
32	18	6–8	84	64	6–8	100	87	4–5*
40	29	4–6	84	71	4–6	102	82	6–8
42	31	4–6	86	46	12–14	102	89	4–5*
44	33	4–5*	86	66	6–8	104	84	6–8
46	16	12–14	86	73	4–6	104	91	4–5*
46	25	8–10	88	48	12–14	106	86	6–8
46	35	4–5*	88	68	6–8	106	93	4–5*
48	26	8–11*	88	75	4–6	108	88	6–8
48	32	6–8	90	50	12–14	108	95	4–5*
48	37	4–5*	90	70	6–8	110	90	6–8
50	39	4–5*	90	77	4–6	110	97	4–5*
52	41	4–5*	92	52	12–14	112	92	6–8
54	43	4–5*	92	72	6–8	112	99	4–5*
56	45	4–5*	92	79	4–5*	114	94	6–8
58	47	4–5*	94	54	12–14	114	101	4–5*
60	49	4–5*	94	74	6–8	116	96	6–8
62	51	4–5*	94	81	4–5*	116	103	4–5*
72	53	6–8	96	31	22–24	118	98	6–8
74	55	6–8	96	76	6–8	118	105	4–5*
74	61	4–6	96	83	4–5*	120	100	6–8
76	57	6–8	98	19	28–30	120	107	4–5*
76	63	4–6	98	31	23–24	122	102	6–8
78	59	6–8	98	32	22–24	122	109	4–5*
78	65	4–6	98	78	6–8	124	104	6–8
						124	111	4–5*

Acknowledgment. We thank an anonymous reviewer for providing a simpler proof and a more general statement of Theorem 1, in addition to numerous comments on style and organization.

REFERENCES

[1] H. J. HELGERT AND R. D. STINOFF, *Minimum-distance bounds for binary linear codes*, IEEE Trans. Inform. Theory, IT-19 (1973), pp. 344–356.

[2] P. PIRET, *Convolutional codes and irreducible ideals*, Philips Research Reports, 27 (1972), pp. 257–271.

[3] ———, *Structure and construction of cyclic convolutional codes*, IEEE Trans. Inform. Theory, IT-22 (1976), pp. 147-155.

SIAM. J. ALG. DISC. METH.
Vol. 2, No. 2, June 1981

0196-5212/81/0202-0006 $01.00/0

MAXIMUM SEMIORDERS IN INTERVAL ORDERS*

PETER C. FISHBURN†

Abstract. Let $s(n)$ be the largest integer such that every n-point interval order includes an $s(n)$-point semiorder. Equivalently, $s(n)$ is the largest integer such that every n-point interval graph includes an $s(n)$-point unit interval graph. Although $s(n-1)=s(n)=n/2+1$ for even n from 4 to 14, this pattern does not persist since $s(17)=9$. In addition, $s(n)>n/\log_2 n$ for $n\geqq 3$, and $s(n)/n\to 0$. It is conjectured that $s(n)\ (\log_2 n)/n\to c$ for some c in [1, 3].

1. Introduction. Throughout this paper we shall assume that $\prec$ is an asymmetric and transitive binary relation on a nonempty finite set X, with symmetric complement $\sim$, so that $x\sim y$ iff neither $x\prec y$ nor $y\prec x$. A *chain* in $(X, \prec)$ is a subset of X linearly ordered by $\prec$, and an *antichain* in $(X, \prec)$ is a subset of X whose points all stand in the relation $\sim$ to one another. We shall say that $\prec$ on X, or $(X, \prec)$, is an *interval order* if

$$\forall x, y, z, w\in X:\quad (x\prec y \text{ and } z\prec w)\Rightarrow(x\prec w \text{ or } z\prec y),$$

and a *semiorder* if it is an interval order such that

$$\forall x, y, z, w\in X:\quad (x\prec y \text{ and } y\prec z)\Rightarrow(x\prec w \text{ or } w\prec z).$$

The latter property is sometimes referred to as semitransitivity [2].

Semiorder and interval order specializations of partial orders were introduced, respectively, by Luce [10] and Fishburn [3], [4], and have been examined extensively by others [1], [7], [8], [12], [15], [16], [17], [18], [19]. The present paper considers a question that is not unlike the question of the largest integer $t(n)$ such that every tournament on n points includes a transitive subtournament on at least $t(n)$ points [11], [13]. In particular, we shall consider the largest integer $s(n)$ such that every interval order on n points includes a semiorder (semitransitive interval order) on at least $s(n)$ points. Formally, $s(n)$ is the largest integer such that, for every interval order $(X, \prec)$ with $|X|=n$, there exists a semiorder $(X^*, \prec^*)$ such that $X^*\subseteq X$, $\prec^*=\prec\cap(X^*\times X^*)$ and $|X^*|\geqq s(n)$.

A graph-theoretic description of s that does not refer directly to transitivity can be developed through the following representation theorem. Let $\mathcal{J}$ be the set of all nondegenerate closed real intervals that have finite lengths.

THEOREM 1. ([4], [16], [17]). *$(X, \prec)$ is an interval order iff there is a mapping I: $X\to\mathcal{J}$ such that*

$$\forall x, y\in X:\quad x\prec y \quad \textit{iff} \quad \sup I(x)<\inf I(y);$$

and, when this representation holds, $(X, \prec)$ is a semiorder iff no $I(x)$ intersects each of three pairwise disjoint $I(y)$. Moreover, $(X, \prec)$ is a semiorder iff there is such a mapping for which all intervals have unit length.

When $(X, \prec)$ is an interval order, we shall say that a four-point subset of X (or the interval configuration for these four points) is a *Q-set* iff three of the four points form a chain and the fourth point stands in the relation $\sim$ to the other three. By Theorem 1 and our earlier definitions, an interval order is a semiorder if it includes no Q-set, and $s(n)$ is the largest integer such that every interval order on n points includes an $s(n)$-point subset that includes no Q-set.

* Received by the editors July 11, 1980, and in final form October 15, 1980.
† Bell Telephone Laboratories, Murray Hill, New Jersey 07974.

In terms of graphs, we shall set aside the usual convention and consider graphs that contain all loops, since this is convenient with respect to $\sim$. With this modification, we shall say that $(X, \sim)$ is an *interval graph* [5], [6], [9] iff there exists $I: X \to \mathcal{I}$ such that

$$\forall x, y \in X: \quad x \sim y \quad \text{iff} \quad I(x) \cap I(y) \neq \varnothing,$$

and that $(X, \sim)$ is a *unit interval graph* [14] if there is such an I for which all intervals have unit length. By Theorem 1, $(X, \sim)$ is an interval graph when $(X, \prec)$ is an interval order, and $(X, \sim)$ is a unit interval graph when $(X, \prec)$ is a semiorder. When $(X, \sim)$ is an interval graph (unit interval graph), there may be several interval orders (semiorders) $(X, \prec)$ for which $\sim$ is the symmetric complement of $\prec$.

This brings us to the alternative description of s alluded to above: $s(n)$ is the largest integer such that every n-point interval graph includes an $s(n)$-point unit interval graph. Alternatively, every n-point interval graph includes an $s(n)$-point induced subgraph that does not include any K_{13} subgraph [14] (with loops), and $s(n)$ is the largest integer for which this is true.

Clearly, $s(n) = n$ for $n \leqq 3$. For even n from 4 to 14, we argue that $s(n-1) = s(n) = n/2+1$. If this pattern persisted then it would give $s(17) = 10$, but in fact we shall see that $s(17) = 9$. Since s cannot decrease in n, $s(15)$ and $s(16)$ are in $\{8, 9\}$. It has been determined that $s(15) = s(16) = 9$, but my proof of this is long and will not be given here.

As n increases, it appears that the rate of increase of s diminishes gradually. We shall prove that $s(n) \to \infty$ while $s(n)/n \to 0$ as $n \to \infty$. In particular, we shall prove that $s(n) > n/\log_2 n$ for $n \geqq 3$, while

$$s(2^k(k+4)) \leqq 3(2^k) \quad \text{for } k = 0, 1, \cdots.$$

It follows from the latter result that $s(n) < 7n/\log_2 n$ for all $n \geqq 2$, and that $s(n) \times (\log_2 n)/n$ for $n \in \{2^k(k+4): k = 0, 1, \cdots\}$ cannot converge to a value exceeding 3 as $n \to \infty$. Hence, it is tempting to conjecture that $s(n)(\log_2 n)/n \to c$ for some $c \in [1, 3]$, but this remains open.

Two simple observations will be used without special mention in the proofs of the foregoing results. First, the largest semiorder included in an interval order must contain at least as many points as are contained in any two disjoint antichains. (A Q-set cannot be formed from the points in two antichains.) Second, if $X_1, \cdots, X_N$ are subsets of X for which $X_i \prec X_{i+1}$ (i.e., $x_i \prec x_{i+1}$ for all $x_i \in X_i$ and $x_{i+1} \in X_{i+1}$) for $i = 1, \cdots, N-1$, then the largest semiorder in the interval order $(X, \prec)$ must contain at least $\sum_i s'(X_i)$ points, where $s'(X_i)$ is the largest semiorder in X_i.

2. Small n. Our first result gives an upper bound on s that turns out to be tight for small n, but only for small n.

LEMMA 1. *$s(n-1) \leqq s(n) \leqq n/2+1$ for even $n \geqq 4$.*

Proof. For even $n \geqq 4$, let A_n be an n-point interval order consisting of an $(n/2+1)$-point chain plus $n/2-1$ points that bear $\sim$ to each other and to every point in the chain. (A_4 is a Q-set.) This is pictured in Fig. 1(a), where intervals are arranged vertically as well as horizontally for visual convenience. In the figure, $x \prec y$ iff $I(x)$ lies wholly to the left of $I(y)$, and the integer written immediately above an interval is the number of points in X that are mapped into that interval.

The largest semiorders in A_n are clearly the $(n/2+1)$-point chain and the $(n/2+1)$ points in two antichains that include the $n/2-1$ points for the long interval along with two points in the chain. Hence $s(n) \leqq n/2+1$ according to the definition of $s(n)$. Since s does not decrease in n, $s(n-1) \leqq s(n)$. □

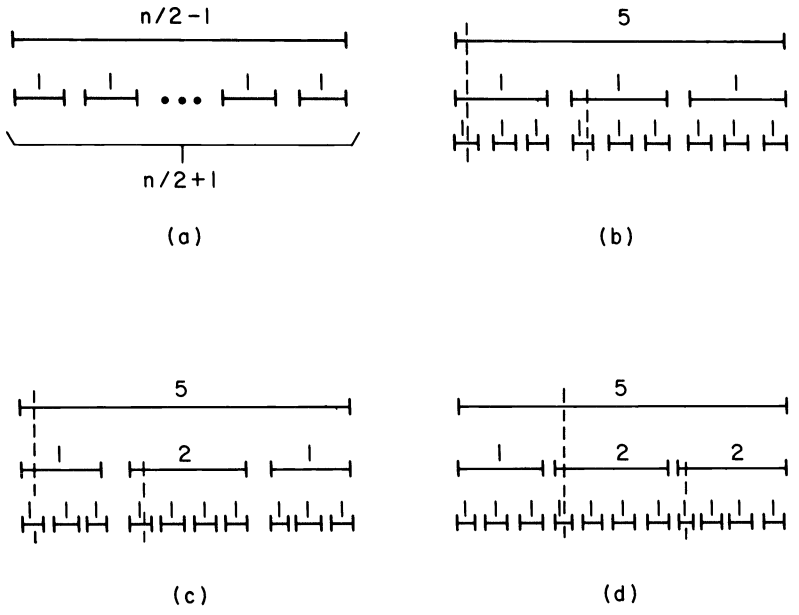


FIG. 1

Similar reasoning based on Figs. 1(b) through 1(d) shows that $s(17) \leqq 9$, $s(19) \leqq 10$ and $s(21) \leqq 11$. The dashed lines in these figures identify two disjoint antichains that maximize the number of points in two antichains of the illustrated interval orders.

THEOREM 2. *$s(n-1) = s(n) = n/2+1$ for even n from 4 to 14 inclusive.*

Proof. In view of Lemma 1, it suffices to show that $s(m) \geqq (m+3)/2$ for odd m from 3 to 13. Clearly $s(3) = 3$, and I shall leave the proofs for $m = 5, 7, 9$ to the reader.

To prove that $s(11) \geqq 7$, let $(X, <) = (\{a_1 \cdots, a_{11}\}, <)$ be a generic eleven-point interval order with interval assignment I as in Theorem 1. With $I_i = I(a_i)$, $I_i^- = \inf I_i$ and $I_i^+ = \sup I_i$, arrange subscripts so that $I_1^- \leqq I_2^- \leqq \cdots \leqq I_{11}^-$. Suppose first that $I_i^+ < I_7^-$ for at least three $i \leqq 6$. Then, since all I_i for $i \geqq 7$ begin after these three or more end, $(X, <)$ must include a semiorder with at least $s(3)+s(5) = 3+4 = 7$ points.

Suppose next that $I_i^+ < I_7^-$ for two or fewer $i \leqq 6$. Then at least five points in $\{a_1, \cdots, a_7\}$, including a_7, form an antichain. If $\{a_8, \cdots, a_{11}\}$ includes a two-point antichain, then we have two disjoint antichains with at least seven points. Assume henceforth that $\{a_8, \cdots, a_{11}\}$ does not include a two-point antichain, so that it gives the four-point chain $a_8 < a_9 < a_{10} < a_{11}$. Then, if $I_i^+ < I_8^-$ (i.e., $a_i < a_8$) for three or more $i \leqq 7$, we obtain a semiorder with at least $s(3)+4 = 7$ points. Otherwise, if fewer than three I_i^+ are less than I_8^-, then $\{a_1, \cdots, a_8\}$ includes an antichain with at least six points. Such an antichain, plus any other point, gives a semiorder with at least seven points. This completes the proof that $s(11) \geqq 7$.

Henceforth, let $(\{a_1, \cdots, a_{13}\}, <)$ be a 13-point interval order with $I_1^- \leqq \cdots \leqq I_{13}^-$. We consider four exhaustive cases as follows to prove that $s(13) \geqq 8$.

Case 1. $I_i^+ < I_8^-$ for at least five $i \leqq 7$. Then there is a semiorder with at least $s(5)+s(6) = 8$ points.

Case 2. $I_i^+ < I_8^-$ for two or fewer $i \leqq 7$. Then $\{a_1, \cdots, a_8\}$ includes a six-point antichain. If $\{a_9, \cdots, a_{13}\}$ has a two-point antichain then we are done, so suppose henceforth that a_9 through a_{13} form a five-point chain. If $I_i^+ < I_9^-$ for three or more $i \leqq 8$, we get a semiorder with at least $s(3)+5 = 8$ points; if $I_i^+ < I_9^-$ for less than

three $i \leqq 8$, then $\{a_1, \cdots, a_9\}$ includes a seven-point antichain and hence an eight-point semiorder.

Case 3. $I_i^+ < I_8^-$ for exactly three $i \leqq 7$, so that $\{a_1, \cdots, a_8\}$ includes a five-point antichain that contains a_8. In this case we shall assume that there is no eight-point semiorder and derive a contradiction. Thus, assume henceforth that $\{a_9, \cdots, a_{13}\}$ includes no three-point antichain. Assume also that $I_i^+ < I_9^-$ for at most four $i \leqq 8$, since otherwise we get a semiorder with at least $s(5)+s(5)=8$ points.

Suppose that exactly three $i \leqq 8$ have $I_i^+ < I_9^-$. Then $\{a_1, \cdots, a_9\}$ includes a six-point antichain. Hence, to preclude an eight-point semiorder, we require $a_{10} < a_{11} < a_{12} < a_{13}$. Given this four-point chain: if $I_i^+ < I_{10}^-$ for at least five $i \leqq 9$, we get a semiorder with at least $s(5)+4=8$ points; if fewer than four $i \leqq 9$ have $I_i^+ < I_{10}^-$, then $\{a_1, \cdots, a_{10}\}$ includes a seven-point antichain and hence an eight-point semiorder; and if exactly four $i \leqq 9$ have $I_i^+ < I_{10}^-$, then $\{a_1, \cdots, a_{10}\}$ includes a six-point antichain, so that the four a_i with $I_i^+ < I_{10}^-$ must form a chain to prevent an eight-point semiorder, in which case these four a_i plus a_{10} through a_{13} form an eight-point chain.

Hence, if there is no eight-point semiorder, exactly four $i \leqq 8$ have $I_i^+ < I_9^-$. These four consist of the set A of the three $i \leqq 7$ for which $I_i^+ < I_8^-$, plus another a_i with $i \leqq 7$, which we denote as a_j. (If the fourth point were a_8 then $A \cup \{a_8\}$ would be a semiorder, which along with a four-point semiorder from $\{a_9, \cdots, a_{13}\}$ would yield an eight-point semiorder overall.) Let $B = \{a_1, \cdots, a_7\} \backslash (A \cup \{a_j\})$ comprise the other three points from $\{a_1, \cdots, a_7\}$, each of which bears the relationship $\sim$ to a_8 and a_9. To prevent an eight-point semiorder, it is easily seen that $A \cup \{a_j\}$ must be a Q-set, and all three points in B must bear $\sim$ to a point in A, say a_k, whose interval extends farthest right. Hence $B \cup \{a_j, a_k\}$ is a five-point antichain, which requires that there be no three-point antichain in $\{a_8, a_9, \cdots, a_{13}\}$. Fig. 2 indicates the interval picture that applies at this time in our analysis of Case 3.

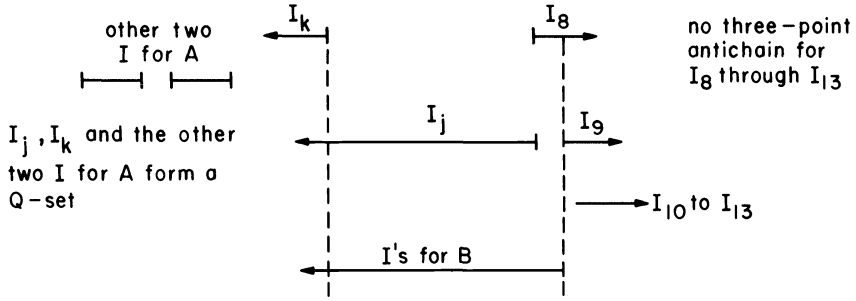


FIG. 2

With respect to Fig. 2, suppose first that $I_9^+ < I_{10}^-$, or $a_9 < a_{10}$. If $\{a_{10}, \cdots, a_{13}\}$ form a Q-set, then I_8 can intersect at most I_{10} of I_{10} through I_{13}, and we get an eight-point semiorder for $A \cup \{a_8, a_9\} \cup \{a_{11}, a_{12}, a_{13}\}$; if $\{a_{10}, \cdots, a_{13}\}$ does not form a Q-set, then $A \cup \{a_9\} \cup \{a_{10}, \cdots, a_{13}\}$ forms an eight-point semiorder. Suppose next that $a_9 \sim a_{10}$, so that I_9 and I_{10} intersect. Then I_8 must end before I_{10} begins: if $a_9 \sim a_{11}$ then $a_{10} < a_{11}$ and $A \cup \{a_8, a_{10}, a_{11}, a_{12}, a_{13}\}$ yields an eight-point semiorder; if $a_9 < a_{11}$ then $A \cup \{a_8, a_9, a_{11}, a_{12}, a_{13}\}$ forms an eight-point semiorder. Since this exhausts the possibilities, there must in fact be an eight-point semiorder in $(\{a_1, \cdots, a_{13}\}, <)$.

Case 4. $I_i^+ < I_8^-$ for exactly four $i \leqq 7$. Let C be these four a_i, and $D = \{a_1, \cdots, a_7\} \backslash C$. If C is not a Q-set, then C plus four from $\{a_8, \cdots, a_{13}\}$ gives an eight-point semiorder. Assume henceforth that C is a Q-set, with a_j and a_k its points whose

intervals extend farthest to the right. If at least one point in D does not bear $\sim$ to both a_j and a_k, then this point, along with a_8 through a_{13}, has a five-point semiorder (since $s(7)=5$), which combines with a three-point semiorder from C to yield an eight-point semiorder. Assume that all three points in D bear $\sim$ to both a_j and a_k, so that these five points form an antichain. Then, if $\{a_8, \cdots, a_{13}\}$ includes a three-point antichain, we get an eight-point semiorder. Otherwise, a picture similar to Fig. 2 applies, and an argument like that used in the preceding paragraph shows that there must be an eight-point semiorder. □

Our next result shows that the pattern set in Theorem 2 breaks down by $n=17$.

THEOREM 3. $s(17)=9$.

Proof. Since $s(17)\leqq 9$ by Fig. 1(b), we need only show that $s(17)\geqq 9$. Proceeding as in the proof of Theorem 2 with $I_1^- \leqq \cdots \leqq I_{17}^-$, let k be the number of $i\leqq 8$ for which $I_i^+ < I_9^-$. If $k\geqq 3$, then a three-point semiorder from these k plus a six-point semiorder from $\{a_9, \cdots, a_{17}\}$ yields a nine-point semiorder overall. If $k\leqq 1$, then $\{a_1, \cdots, a_9\}$ is a semiorder (one or two antichains). If $k=2$, then $\{a_1, \cdots, a_9\}$ includes a seven-point antichain. Then, if $\{a_{10}, \cdots, a_{17}\}$ includes a two-point antichain, we are done. Otherwise, $\{a_{10}, \cdots, a_{17}\}$ forms an eight-point chain, which along with the special $k=2$ elements from $\{a_1, \cdots, a_8\}$ yields a ten-point semiorder. □

3. Large *n*. Our next result gives a lower bound on s which shows that $s(n)$ is unbounded. We then consider upper bounds.

THEOREM 4. $s(n) > n/\log n$, *for all* $n\geqq 3$.

Remark. Here and later, all logarithms are to base 2.

Proof. By Theorem 2, the inequality on $s(n)$ in Theorem 4 holds for small n. For larger n, we shall presume that the result holds for $n' < n$ and prove that every interval order on n points includes a semiorder on at least $n/\log n$ points. Throughout the proof, $h(r)=r/\log r$ for real $r>0$, and $g(r)$ is the smallest integer as great as $h(r)$.

Let $(X, <)$ be an interval order on n points, with interval representation I as in Theorem 1. If $(X, <)$ includes two antichains with at least $g(n)$ points, then we are done. Assume henceforth that no two antichains have more than $g(n)-1$ points.

For each real r, let

$$f_1(r)=|\{x: \sup I(x) < r\}|,$$
$$f_2(r)=|\{x: \inf I(x) > r\}|,$$

let r_i be an r that minimizes $f_i(r)$ subject to

$$f_i(r)\geqq \frac{n+2-2g(n)}{3}$$

and let $n_i = f(r_i)$ for $i=1, 2$. The last jump in $f_1(r)$ before r_1 (as r increases) and the last jump in $f_2(r)$ before r_2 (as r decreases), cannot involve more than $g(n)-1$ points by the two-antichains restriction. Therefore

$$n_1+n_2 < \frac{2[n-2-2g(n)]}{3}+g(n)-1=\frac{2n+1-g(n)}{3}.$$

Let n_3 be the number of points in X whose intervals lie strictly between r_1 and r_2. Since no more than $g(n)-1$ points have intervals that contain r_1 or r_2, $n_1+n_2+n_3\geqq$

$n+1-g(n)$, and therefore

$$\begin{aligned} n_3 &\geqq n+1-g(n)-(n_1+n_2) \\ &> n+1-g(n)-\frac{2n+1-g(n)}{3} \\ &= \frac{n+2-2g(n)}{3}. \end{aligned}$$

Thus $(X, \prec)$ includes a semiorder on at least $g(n_1)+g(n_2)+g(n_3)$ points, with

$$n_i \geqq \frac{n+2-2g(n)}{3} \quad \text{for } i=1,2,3,$$

$$n_1+n_2+n_3 \geqq n+1-g(n).$$

By definition, $\sum g(n_i) \geqq \sum h(n_i)$. Since $h(r)$ is concave increasing for $r>4$, $\sum h(n_i)$ is minimized subject to the constraints on n_i, when n_1 and n_2 are made as small as possible and $n_3 = n+1-g(n)-(n_1+n_2)$. Since $-g(n) > -(n/\log n+1)$,

$$n_i \geqq \frac{n+2-2\left(\frac{n}{\log n}+1\right)}{3} = \frac{n-\frac{2n}{\log n}}{3}$$

for $i=1$, 2, and $n_3 \geqq n+1-(n/\log n+1)-(n_1+n_2)$. This lower bound on n_3 equals $[n+n/\log n]/3$ when $n_1=n_2=[n-2n/\log n]/3$. It follows that

$$\sum h(n_i) \geqq 2h\left(\frac{n-\frac{2n}{\log n}}{3}\right)+h\left(\frac{n+\frac{n}{\log n}}{3}\right).$$

Hence, if the right-hand side of this inequality is greater than $h(n)$, then $\sum g(n_i) > n/\log n$ and the proof is complete.

Thus, it remains to show that

$$\frac{\frac{2(n-2n/\log n)}{3}}{\log\left(\frac{n-2n/\log n}{3}\right)}+\frac{\frac{(n+n/\log n)}{3}}{\log\left(\frac{n+n/\log n}{3}\right)} > \frac{n}{\log n},$$

say, for $n \geqq 32$, since our previous analysis shows that Theorem 4 is true for smaller values of n. After cancellation and rearrangement, the preceding inequality can be expressed as

$$\begin{aligned} &3\,(\log 3-1)\log n+(\log n)\log\left\{\frac{(\log n)^3}{(\log n)^3-[(\log n)^2-2\log n-4]}\right\} \\ &\quad > 3\,(\log 3-1)\log 3+(3\log 3+1)\log\left(\frac{\log n}{\log n-2}\right) \\ &\qquad -\left[\log\left(1+\frac{1}{\log n}\right)\right]\left[3\log\left(\frac{\log n}{\log n-2}\right)+(3\log 3-4)\right]. \end{aligned}$$

For $n \geqq 32$, the left-hand side exceeds its first term and the right-hand side is less than

the sum of its first two terms. Therefore the inequality holds if

$$\log n > \log 3 + \frac{3 \log 3 + 1}{3 (\log 3 - 1)} \log \left(\frac{\log n}{\log n - 2} \right).$$

This is true when $n = 32$ and, since its left-hand side increases in n while its right-hand side decreases in n, it is true for all $n \geqq 32$. □

We now develop an upper bound on $s(n)$ that is considerably sharper than the bound in Lemma 1, for large n. This is done by constructing a symmetric hierarchical series C_0, C_1, C_2, $\cdots$ of successively larger interval orders such that the largest semiorder in each C_k is realized (among other ways) both by a chain and by two antichains. The first order in the series, C_0, consists of an $(n/2+1)$-point chain and $n/2-1$ other points that bear $\sim$ to everything else. That is, C_0 is an A_n as described in Fig. 1(a).

Given C_k, the next order, C_{k+1}, consists of two copies of C_k, one of which lies completely to the left of the other (one copy $\prec$ other copy), plus δ_k other points that bear $\sim$ to everything else, such that

$$\begin{aligned} \delta_k + 2 \text{ (maximum number of points in an antichain of } C_k) \\ = \text{maximum number of points in two antichains of } C_{k+1} \\ = \text{number of points on the maximum chain of } C_{k+1} \\ = 2 \text{ (number of points in the maximum chain of } C_k). \end{aligned}$$

Fig. 3 illustrates the construction when C_0 is the Q-set A_4. The figure shows C_3, as built up from two copies of C_2, four copies of C_1 and eight copies of C_0. It should be apparent

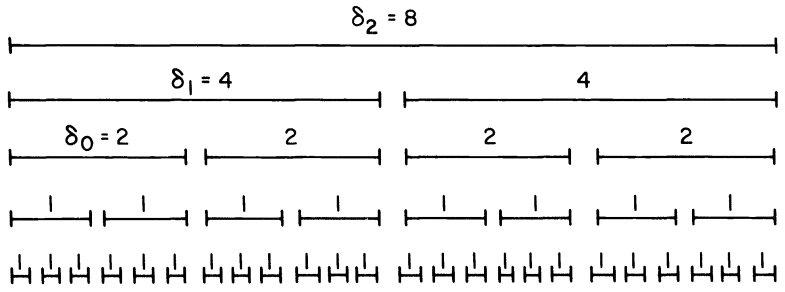


FIG. 3. C_3 *based on* A_4.

from the figure that if the δ_2 points are used for a semiorder in C_3, then the largest such semiorder will consist of two antichains, with a total of $8+2(4+2+1+1) = 24$ points. On the other hand, if none of the δ_2 points are used for a semiorder in C_3, then the largest such semiorder has twice as many points as the largest semiorder in C_2, namely $2(12) = 24$. The latter maximum semiorder in C_3 can be realized in several ways, one of which is the 24-point chain consisting of the shortest intervals on the figure. Hence $s(56) \leqq 24$.

For each C_k let

α_k = number of points in C_k,

β_k = number of points in a maximum antichain in C_k,

γ_k = number of points in a maximum semiorder in C_k,

δ_k = number of points added to the two copies of C_k to give C_{k+1}.

The equalities in the preceding paragraph give

$$\delta_k + 2\beta_k = \gamma_{k+1} = 2\gamma_k.$$

According to this and the construction,

$$\alpha_{k+1} = 2\alpha_k + \delta_k,$$

$$\beta_{k+1} = \beta_k + \delta_k,$$

$$\gamma_{k+1} = 2\gamma_k = 2\beta_k + \delta_k,$$

$$\delta_{k+1} = 2(\gamma_{k+1} - \beta_{k+1}) = 2\beta_k.$$

The next theorem shows what we can conclude from this recursive scheme when $C_0 = A_4$.

THEOREM 5. $s(2^k(k+4)) \leqq 3(2^k)$ *for* $k = 0, 1, \cdots$. *The resultant upper bound on* $s(n)(\log n)/n$ *for* $n \in \{2^k(k+4): k = 0, 1, \cdots\}$ *approaches* 3.

Proof. Given $C_0 = A_4$, $(a_0, \beta_0, \gamma_0, \delta_0) = (4, 2, 3, 2)$. It then follows easily from the recursive scheme that $\gamma_k = 3(2^k)$ and $\alpha_k = 2^k(k+4)$. Hence, by the definitions of α_k, γ_k and s, $s(2^k(k+4)) \leqq 3(2^k)$. With $n = 2^k(k+4)$, $\log n = k + \log(k+4)$, and therefore

$$s(n)\frac{\log n}{n} \leqq \frac{3[k + \log(k+4)]}{k+4}.$$

The right-hand side of this inequality approaches 3 as k gets large. □

The following corollary of Theorem 5 provides an upper bound on $s(n)$ for all $n > 1$.

COROLLARY 1. *For any* $\delta > 0$, $s(n) < (6+\delta)n/\log n$ *for* n *sufficiently large, and* $s(n) < 7n/\log n$ *for all* $n \geqq 2$.

Proof. With $\alpha_k = 2^k(k+4)$, $\alpha_k \leqq n \leqq \alpha_{k+1}$ gives $s(n) \leqq s(\alpha_{k+1}) \leqq 3(2^{k+1}) = 6(2^k) < 7(2^k)(k+4)/[k + \log(k+4)] = 7\alpha_k/\log \alpha_k \leqq 7n/\log n$, so $s(n) < 7n/\log n$ for all $n \geqq \alpha_0 = 4$. The same bound on $s(n)$ holds for $n \in \{2, 3\}$. When 7 is replaced by $6+\delta$ in the preceding series of inequalities, we get $6(2^k) < (6+\delta)2^k(k+4)/[k + \log(k+4)]$ for sufficiently large k, and therefore $s(n) < (6+\delta)n/\log n$ for sufficiently large n. □

We used $C_0 = A_4$ for Theorem 5 since the successive terms in the recursive scheme are easiest to compute in this case. However, similar conclusions are obtained when we start with other A_m. In particular, the bound obtained on $s(n)(\log n)/n$ always converges to 3 regardless of which A_m is used for C_0. To see this, we note without proof that the recursive scheme gives

$$\gamma_k = 2^k\gamma_0,$$

$$\beta_k = 2[2^k + (-1)^{k+1}]\frac{\gamma_0}{3} + (-1)^k\beta_0,$$

$$\delta_k = 2[2^k + (-1)^{k+1}]\frac{\gamma_0}{3} + (-1)^k\delta_0,$$

$$\alpha_k = 2^k\alpha_0 + [2^k + (-1)^{k+1}]\frac{\delta_0}{3} + 2[(3k-2)2^{k-1} + (-1)^k]\frac{\gamma_0}{9}$$

$$= 2^k\left[\alpha_0 + \frac{\delta_0}{3} + (3k-2)\frac{\gamma_0}{9}\right] + (-1)^{k+1}\left(\frac{\delta_0}{3} - \frac{2\gamma_0}{9}\right).$$

With $n = \alpha_k$, $s(n)(\log n)/n \leqq \gamma_k(\log \alpha_k)/\alpha_k$, and it follows from the equations just given that $\gamma_k(\log \alpha_k)/\alpha_k \to 3$.

4. Discussion. The main open questions for large n that arise from the preceding analysis are whether $s(n)(\log n)/n$ converges, and, if so, to what value in [1, 3]. My best guess is that $s(n)(\log n)/n \to 3$.

As noted earlier, it is known that $s(15) = s(16) = 9$, but a proof of this (available from the author) has been omitted due to its length. Thus, $s(n)$ is known precisely for n from 1 through 17: the series of $s(n)$ values is 1, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9, 9. The question of whether the number of successive n for which $s(n) = m + 1$ is as great as the number of successive n for which $s(n) = m$ is open. It would also be interesting to know if the bound in Theorem 5 were exact, i.e., if $s(2^k(k+4)) = 3(2^k)$ for $k = 0, 1, 2, \cdots$. Our earlier results show only that equality holds for $k \in \{0, 1\}$.

REFERENCES

[1] K. P. BOGART, I. RABINOVITCH AND W. T. TROTTER, *A bound on the dimension of interval orders*, J. Combin. Theory Ser. A, 21(1976), pp. 319–328.

[2] J. S. CHIPMAN, *Consumption theory without transitive indifference*, Preferences, Utility, and Demand, J. S. Chipman, L. Hurwicz, M. K. Richter and H. F. Sonnenschein, eds., Harcourt Brace Jovanovich, New York, 1971, pp. 224–253.

[3] P. C. FISHBURN, *Intransitive indifference with unequal indifference intervals*, J. Math. Psych., 7(1970), pp. 144–149.

[4] ———, *Utility Theory for Deçision Making*, John Wiley, New York, 1970.

[5] D. R. FULKERSON AND O. A. GROSS, *Incidence matrices and interval graphs*, Pacific J. Math., 15(1965), pp. 835–855.

[6] P. C. GILMORE AND A. J. HOFFMAN, *A characterization of comparability graphs and of interval graphs*, Canad. J. Math., 16(1964), pp. 539–548.

[7] T. L. GREENOUGH AND K. P. BOGART, *The representation and enumeration of interval orders*, Discrete Math., to appear.

[8] K. H. KIM AND F. W. ROUSH, *Enumeration of isomorphism classes of semiorders*, J. Combinatorics, Information System Science, 3(1978), pp. 58–61.

[9] 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.

[10] R. D. LUCE, *Semiorders and a theory of utility discrimination*, Econometrica, 24(1956), pp. 178–191.

[11] J. W. MOON, *Topics on Tournaments*, Holt, Rinehart and Winston, New York, 1968.

[12] I. RABINOVITCH, *The dimension of semiorders*, J. Combin. Theory Ser. A, 25(1978), pp. 50–61.

[13] K. B. REID AND E. T. PARKER, *Disproof of a conjecture of Erdös and Moser on tournaments*, J. Combin. Theory, 9(1970), pp. 225–238.

[14] F. S. ROBERTS, *Indifference graphs*, in Proof Techniques in Graph Theory, F. Harary, ed., Academic Press, New York, 1969, pp. 139–146.

[15] ———, *Homogeneous families of semiorders and the theory of probabilistic consistency*, J. Math. Psych., 8(1971), pp. 248–263.

[16] D. SCOTT, *Measurement structures and linear inequalities*, J. Math. Psych., 1(1964), pp. 233–247.

[17] D. SCOTT AND P. SUPPES, *Foundational aspects of theories of measurement*, J. Symbolic Logic, 23(1958), pp. 113–128.

[18] W. T. TROTTER AND K. P. BOGART, *Maximal dimensional partially ordered sets* III: *a characterization of Hiraguchi's inequality for interval dimension*, Discrete Math., 15(1976), pp. 389–400.

[19] ———, *On the complexity of posets*, Discrete Math., 16(1976), pp. 71–82.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 2, June 1981

0196-5212/81/1202-0007 $01.00/0

EIGENVECTORS OF A TOEPLITZ MATRIX: DISCRETE VERSION OF THE PROLATE SPHEROIDAL WAVE FUNCTIONS*

F. ALBERTO GRÜNBAUM†

Abstract. The discrete Fourier transform leads one, in a natural way, to consider the extent to which a function in Z_N and its transform can both be sharply concentrated. This requires the study of a Toeplitz matrix and its eigenvalues and eigenvectors. For the case at hand this can be done successfully.

Integral operators of convolution type which act on $L^2([-T, T])$ and have a kernel given by

$$K(t, s) = R(t-s), \qquad -T \leqq t, s \leqq T$$

arise in numerous applications. The discrete version of these operators is given by a matrix

$$K_n(i, j) = r(i-j), \qquad 1 \leqq i, j \leqq n$$

acting on the space of sequences $(x_1, \cdots, x_n)$.

The special nature of these operators suggests that problems involving them should be amenable to a simplified treatment. This is certainly true for the problem of solving linear equations or finding the inverse of these operators; starting with the work of Levinson, Szego and Krein one has very efficient ways to deal with this problem, requiring $O(n \log^2 n)$ instead of the usual n^3 operations. For a nice account of this topic see [6] as well as [11], [12].

Another problem where simplifications could be expected is that of computing eigenvectors and eigenvalues of the operator K, but here the situation is quite different.

If the interval $[-T, T]$ is replaced by $(-\infty, \infty)$, or the set $(1, \cdots, n)$ by the set of all integers or, much in the same spirit, if K_n is "cyclic" or "circulant," the problem is trivial since then the operator K is a simple function of the "shift" operator and thus shares its complete set of eigenfunctions. Except in these simple cases we know of no general method to exploit the Toeplitz nature of K in connection with the eigenvector-eigenvalue problem.

A notable exception is contained in a very detailed study done by Slepian, Landau and Pollak in connection with "prolate spheroidal functions" and the uncertainty principle; see [1], [2] and [3]. They consider the kernel

$$r(\xi) = \frac{\sin \Omega\xi}{\Omega\xi},$$

acting as a convolution integral operator $L^2[-T, T]$; they observe that a second order differential operator can be found which commutes with K, and since both operators have a simple pure point spectrum they must have the same eigenfunctions. In this fashion the problem has been reduced to a much more manageable one.

Notice that $r(\xi)$ is the Fourier transform of an interval symmetrically placed around the origin, and $[-T, T]$ is of the same type.

One could expect that this is the "generic" case, but this is far from true. Indeed, Morrison [13] proved that the only convolution kernels which commute with

* Received by the editors August 8, 1979, and in revised form September 22, 1980.
† Department of Mathematics, University of California, Berkeley, California 94720.

a second order self-adjoint differential operator are given by

$$r(\xi) = \frac{b}{c} \frac{\sin c\xi}{\sin b\xi},$$

for arbitrary complex constants b, c. Morrison's result was never published in full, but it has been quoted in [14]. I learned of it through the kindness of a referee.

Notice that if $r(\xi)$ is required to be the Fourier transform of an L^2 function of compact support, we get $b = 0$ and are back in the case mentioned earlier; so much for the negative side.

On the positive side this situation has been shown by Slepian [4] to hold in the higher dimensional case when intervals are replaced by balls centered at the origin, and more recently Slepian [5] extended this to the case when the interval $[-T, T]$ is replaced by the set of integers $\{-N, \cdots, N\}$ and the other interval is replaced by the interval $-W \leqq \theta \leqq W$ of the unit circle. In this latter case one ends up with a Toeplitz matrix of the form

$$K_{ij} = \frac{\sin 2\pi W(i-j)}{\pi(i-j)}, \qquad -N \leqq i, j \leqq N.$$

One can also (see [5]) exchange the roles of these two "intervals" and end up with an integral operator in $L^2[-W, W]$, with kernel given by

$$K(s, t) = \frac{\sin N\pi(s-t)}{\sin \pi(s-t)}.$$

An important problem for a number of applications would be to go beyond this very restricted situation. Indeed, many "reconstruction" problems can be modeled as follows. The Fourier transform Ff of an unknown function f is known only on the set B and one has the a priori information that f has support in the set A. This is formalized by

$$BFf = g = \text{known},$$

$$Af = f,$$

where A, B denote, by abuse of language, the operators of restriction to the sets A, B. These two equations can be combined into the single equation

$$Ef \equiv BFAf = g,$$

which is best handled by looking at

$$E^*Ef = E^*g.$$

The operator E^*E acts on $L^2(B)$ as a (finite) convolution, with kernel given by the Fourier transform of the characteristic function of the set A. For an example of an "imaging" problem leading to a pair of sets A, B for which the treatment given here has not been possible, see [9].

In the discrete case a natural problem to consider is, thus, the determination of those Toeplitz matrices which allow for a (nontrivial) tridiagonal matrix commuting with them. We have found that, at least in the symmetric case, there is a four-parameter family of Toeplitz matrices with this property: loosely speaking $r(0)$, $r(1)$, $r(2)$, $r(3)$ can be picked arbitrarily and all the other diagonals determined from $r(1)$, $r(2)$, $r(3)$. The proof is rather laborious, due to a number of special cases; it is given in [10].

Notice that in all the examples discussed above we are dealing with an Abelian group G and its dual $\hat{G}$, while A and B are "balls" centered at the identity element of G and $\hat{G}$ respectively.

In this paper we treat the case of the Toeplitz matrix which arises when G is taken to be Z_P, the group of Pth roots of unity. In this case $\hat{G}$ is Z_P again. Although this is really a special case of the result in [10], we feel that being the "discrete-discrete" analogue of the situation studied in [1], [2], [3], [4], [5] it is likely to have a number of applications; therefore giving it without the complications of the more general case discussed in [10] seems worthwhile. One such application is given in [15].

We take A and B to be given as follows:

$$A = \{e^{i2\pi j/P}, |j| \leqq M\}$$

and

$$B = \left\{e^{i2\pi k/P}, |k| \leqq \frac{N}{2}\right\}, \qquad N \text{ even.}$$

In this case the Toeplitz matrix in question has dimensions $N+1 \times N+1$ and is given by

$$K_{ij} = r(i-j), \tag{1}$$

with

$$r(k) = \sum_{l=-M}^{M} e^{2\pi ilk/P} = \frac{\sin(2M+1)(\pi k/P)}{\sin(\pi/P)k} = U_{2M}\left(T_k\left(\cos\frac{\pi}{P}\right)\right). \tag{2}$$

Here T_k and U_{2M} are Chebyshev polynomials of the first and second kind. We notice that these Toeplitz matrices have recently been used in some reconstruction algorithms for X-ray tomography; see [8]. We will find that the situation uncovered by Slepian, Landau and Pollak holds in this instance too.

We show in this paper that the $N+1 \times N+1$ matrix given by (1), (2) commutes with a (essentially unique) tridiagonal matrix. It is advantageous to write this tridiagonal matrix in the form

$$T = D_- A D_+ + B.$$

Here A and B are diagonal matrices and $D_\pm$ stand for the usual difference operators, explicitly,

$$A_{ij} = A_i \delta_{ij}, \quad B_{ij} = b_i \delta_{ij}, \quad D^{\pm}_{ij} = \delta_{i\pm l,j} - \delta_{i,j},$$

with the convention $a_0 = a_{NN+1} = 0$.

One clearly has

$$T = \begin{bmatrix} b_1 - a_1 & a_1 & & & & & \\ a_1 & b_2 - a_2 - a_1 & a_2 & & & & \\ & a_2 & b_3 - a_3 - a_2 & \ddots & & & \\ & & \ddots & \ddots & a_{N-1} & & \\ & & & a_{N-1} & b_N - a_N - a_{N-1} & a_N & \\ & & & & a_N & b_{N+1} - a_N \end{bmatrix}.$$

If K denotes the matrix (1) and we put, as before,

$$K_{ij} = r(i-j),$$

the commutativity condition

$$KT = TK$$

is equivalent to the system of $(N+1)\times(N+1)$ equations

$$(3)\quad \begin{aligned}&(a_i-a_j)(r(i-j+1)-2r(i-j)+r(i-j-1))+(a_i+a_{i-1})(r(i-j)-r(i-j-1))\\&\qquad+(a_j-a_{j-1})(r(i-j+1)-r(i-j))+r(i-j)(b_i-b_j)=0.\end{aligned}$$

We proceed now to solve these equations. Eliminate the b_i's from (3) by setting $j=i-1$ in (3) and $j=i-2$ in (3) to obtain

$$(4)\quad (r(2)-r(1))(a_i-a_{i-2})+r(1)(b_i-b_{i-1})=0$$

and

$$(5)\quad (r(3)-r(2))(a_i-a_{i-3})-(r(2)-r(1))(a_{i-1}-a_{i-2})+r(2)(b_i-b_{i-2})=0.$$

Now replace i by $i-1$ in (4) and add the resulting equation to (4) to obtain

$$(6)\quad (r(2)-r(1))(a_{i-1}-a_{i-3})+(r(2)-r(1))(a_i-a_{i-2})+r(1)(b_i-b_{i-2})=0.$$

Finally multiply (5) by $r(1)$ and (6) by $r(2)$ and subtract to get

$$(7)\quad (a_i-a_{i-3})(r(3)r(1)-r^2(2))+(a_{i-1}-a_{i-2})(r^2(1)-r^2(2))=0.$$

This third order difference equation has for its characteristic equation

$$(8)\quad (\lambda-1)\left[\lambda^2+\lambda\left(1+\frac{r(2)^2-r(1)^2}{r(2)^2-r(1)r(3)}\right)+1\right]=0.$$

Observing that

$$1+\frac{r(2)^2-r(1)^2}{r(2)^2-r(1)r(3)}=2\left(1-2\cos^2\frac{\pi}{P}\right)=-2\cos\frac{2\pi}{P},$$

one can express (8) in the form

$$(\lambda-1)(\lambda-e^{(2\pi i/P)})(\lambda+e^{-(2\pi i/P)})=0,$$

and thus, with the boundary conditions $a_0=a_{N+1}=0$ taken into account, the solution to (7) is given (*except for a multiplicative constant* $\mathbb{C}(P,N)$) as

$$(9)\quad \begin{aligned}a_j&=\sin\frac{2\pi}{P}(N+1)-\sin\frac{2\pi}{P}j-\sin\frac{2\pi}{P}(N+1-j)\\&=2\sin\frac{\pi}{P}(N+1)\left[\cos\frac{\pi}{P}(N+1)-\cos\frac{\pi}{P}(N+1-2j)\right].\end{aligned}$$

To solve for b_i return to (4) and observe that

$$(10)\quad b_j-b_{j-1}=2\frac{r(1)-r(2)}{r(1)}\left[\cos(N+5-2j)\frac{\pi}{P}-\cos(N+1-2j)\frac{\pi}{P}\right]\sin\frac{\pi}{P}(N+1).$$

The vector b_i is determined up to the multiplicative constant $\mathbb{C}(P,N)$ just mentioned and an additive constant which can be adjusted by picking b_1, and we get

$$(11)\quad b_i=b_1+4\frac{r(1)-r(2)}{r(2)}\left[\cos\frac{\pi}{P}N-\cos\frac{\pi}{P}(N+2-2j)\right]\sin(N+1)\frac{\pi}{P}\cos\frac{\pi}{P}.$$

Notice that a_i, b_i as given by (9) and (11) are the (essentially) unique solutions of a very special subsystem of the equations given in (3). Now we proceed to show that they actually satisfy the complete system of equations (3).

Notice that (3) can be rewritten as

$$(3') \qquad \begin{aligned}(r(i-j+1)-r(i-j))(a_i-a_{j-1})+(r(i-j-1)-r(i-j)(a_{i-1}-a_j)\\ +r(i-j)(b_i-b_j)=0,\end{aligned}$$

and thus, omitting a common factor $2\sin(\pi/P)(NN+1)$, we have to check

$$(r(i-j+1)-r(i-j))\left(\cos(N+3-2j)\frac{\pi}{P}-\cos(N+1-2i)\frac{\pi}{P}\right)$$

$$+(r(i-j-1)-r(i-j))\left(\cos(N+1-2j)\frac{\pi}{P}-\cos(N+3-2i)\frac{\pi}{P}\right)$$

$$+2r(i-j)\frac{r(1)-r(2)}{r(1)}\left(\cos(N+2-2j)\frac{\pi}{P}-\cos(N+2-2i)\frac{\pi}{P}\right)\cos\frac{\pi}{P}=0.$$

If $\mathbb{C}$ denotes the factor multiplying $(r(i-j+1)-r(i-j))$ and D the factor multiplying $(r(i-j-1)-r(i-j))$ this can be expressed, after multiplication by $r(1)$, in the form

$$\begin{aligned}&r(1)(r(i-j+1)-r(i-j)\mathbb{C}+r(1)(r(i-j-1)-r(i-j))D+r(i-j)(r(1)-r(2))(\mathbb{C}+D)\\ &\quad=r(1)(\mathbb{C}r(i-j+1)+Dr(i-j-1))-r(2)r(i-j)(\mathbb{C}+D)\\ &\quad=0.\end{aligned}$$

Observing now that

$$\mathbb{C}=-2\sin(N+2-i-j)\frac{\pi}{P}\sin(1+i-j)\frac{\pi}{P},$$

$$D=2\sin(N+2-i-j)\frac{\pi}{P}\sin(1+j-i)\frac{\pi}{P},$$

$$\mathbb{C}+D=-4\sin(N+2-i-j)\frac{\pi}{P}\sin(i-j)\frac{\pi}{P}\cos\frac{\pi}{P},$$

we have to check that

$$-2\frac{\sin(2M+1)(\pi/P)}{\sin(\pi/P)}\left(\sin(2M+1)(i-j+1)\frac{\pi}{P}+\sin(2M+1)(i-j-1)\frac{\pi}{P}\right)$$

$$+4\frac{\sin(2M+1)(2\pi/P)}{\sin(2\pi/P)}\sin(2M+1)(i-j)\frac{\pi}{P}\cos\frac{\pi}{P}=0,$$

which certainly holds.

We conclude with the remark that by choosing the arbitrary constants $\mathbb{C}(P, N)$ and b_1 properly and, letting P and M grow to infinity so that

$$\frac{M}{P}\to W,$$

we obtain in the limit the case discussed by Slepian in [5]. The correct values are

$$\mathbb{C}(P,N)=-\frac{1}{8(\pi/P)^3(N+1)},$$

$$b_1=\frac{N+2}{2}+\cos 2\pi W\frac{N^2}{4},$$

and in this fashion we obtain (in the limit),

$$a_j = \tfrac{1}{2}j(N+1-j)$$

and

$$b_j - a_j - a_{j-1} = \cos 2\pi W\left(\frac{N+1}{2} - j - 1\right)^2, \qquad 1 \leqq j \leqq N+1,$$

as in [5].

REFERENCES

[1] D. SLEPIAN AND H. O. POLLAK, *Prolate spheroidal wave functions, Fourier analysis and uncertainty, I*, Bell System Tech. J., 40 (1961), pp. 43–64.

[2] H. J. LANDAU AND H. O. POLLAK, *Prolate spheroidal wave functions, Fourier analysis and uncertainty, II*, Ibid., 40 (1961), pp. 65–84.

[3] ———, *Prolate spheroidal wave functions, Fourier analysis and uncertainty, III*, Ibid., 41 (1962), pp. 1295–1336.

[4] D. SLEPIAN, *Prolate spheroidal wave functions, Fourier analysis and uncertainty, IV*, Ibid., 43 (1964), pp. 3009–3058.

[5] ———, *Prolate spheroidal wave functions, Fourier analysis and uncertainty, V*, Ibid., 57 (1978), pp. 1371–1430.

[6] T. KAILATH, *Inverse of Toeplitz operators, innovations and orthogonal polynomials*, SIAM Rev., 20 (1978), pp. 106–119.

[7] F. A. GRÜNBAUM, *Second order differential operators commuting with convolution integral operators*, LBL report 9298, 1979.

[8] M. DAVIDSON AND F. A. GRÜNBAUM, *Convolution algorithms for arbitrary projection angles*, IEEE J. Nuclear Science, NS-26 (1979), pp. 2670–2673.

[9] F. A. GRÜNBAUM, *A study of Fourier space methods for "limited angle" image reconstruction*, LBL report 9299, 1979.

[10] ———, *Toeplitz matrices commuting with tridiagonal matrices*, Linear Alg. and Appl., to appear.

[11] F. GUSTAVSON AND D. YUN, *Fast computation for Toeplitz systems, Cauchy-Hermite-Padé approximants, and the extended Euclidean algorithm*, IBM RC 7551, March 1979.

[12] R. BRENT, F. GUSTAVSON AND D. YUN, *Fast computation of Padé approximants and the solution of Toeplitz systems of equations*, to appear.

[13] J. MORRISON, *On the commutation of finite integral operators with difference kernels, and linear selfadjoint differential operators*, Abstract, Notices Amer. Math. Soc., 9 (1962), p. 119.

[14] H. WIDOM, *Asymptotic behavior of eigenvalues of certain integral equations, II*, Arch. Rat. Mech. Anal., 17 (1964), pp. 215–229.

[15] F. A. GRÜNBAUM, *Limited angle reconstruction in tomography*, to appear.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 2, June 1981

0196-5212/81/0202-0008$01.00/0

SET ORDERINGS REQUIRING COSTLIEST ALPHABETIC BINARY TREES*

D. J. KLEITMAN† AND MICHAEL E. SAKS‡

Abstract. It is shown that an ordering of a set with weighted elements which requires the most expensive alphabetic binary tree is a "sawtooth order." For the set $\{e_0, e_1, \cdots, e_t\}$, with the elements indexed from least to greatest weight, this order is $e_0, e_t, e_1, e_{t-1}, \cdots, e_i, e_{t-i}, \cdots$. This result was conjectured by Hwang and leads to an upper bound on the cost of alphabetic binary trees.

1. Introduction. Let E be a finite set whose elements are assigned positive weights. The cost $c(T)$ of a rooted binary tree T whose leaf set is E (such a tree will be called an E-tree) is defined to be $\sum_{e\in E} w(e)\, d_T(e)$, where $d_T(e)$ is the number of arcs in T between the root and e and $w(e)$ is the weight of e. An E-tree T is said to be *alphabetic* with respect to some linear order of E if, in some planar embedding of T, the left-to-right order of the leaves is the given order.

In [6], Huffman described a linear time algorithm for finding the minimum cost E-tree. Hu and Tucker [5] gave an algorithm for finding the minimum cost E-tree which is alphabetic with respect to a given linear order. Garsia and Wachs [1] proposed (and proved correctness of) a variation of the algorithm, and Hu, Kleitman and Tamaki [4] generalized it and provided an elementary proof of correctness.

F. Hwang has raised the following question: given a weighted set E, what linear order maximizes the cost of the minimum cost alphabetic tree. He conjectured that, for a set $\{e_o, e_1, \cdots, e_t\}$ indexed from least to greatest weight, the most expensive order is $e_0, e_t, e_1, e_{t-1}, e_2, e_{t-2}, \cdots$. It is the purpose of this paper to give a proof of this conjecture, thereby characterizing the "worst case" cost of an alphabetic tree.

This problem arose from an effort to make general statements about the cost imposed on the optimal E-tree by restriction to alphabetic trees. Characterizing the "worst case" order gives rise to a bound on the cost of this restriction. The bound implied by the above solution is as follows. If $\{w_1, \cdots, w_m\}$ is the set of weights of a given set, let $H(\{w_1, \cdots, w_m\})$ be the cost of the optimal Huffman (unrestricted) tree on the set. The cost of the optimal alphabetic tree for the worst case ordering is then

$$H\left(\left\{w_i + w_{m-i+1} \mid i = 1, \cdots, \frac{m}{2}\right\}\right) + \sum_{i=1}^{m} w_i$$

for m even, and

$$H\left(\left\{\frac{w_{m+1}}{2}\right\} \cup \left\{w_i + w_{m-i+1} \mid i = 1, \cdots, \frac{m-1}{2}\right\}\right) + \sum_{i=1}^{m} w_i - w_{(m+1)/2}$$

for m odd.

In what follows, capital letters will usually represent a sequence of elements, that is, a set together with a fixed linear order. A small letter will denote individual elements. A comma is used to concatenate sequences and elements: E_1, e_1, E_2, e_2, E_3 is the set $E_1 \cup E_2 \cup E_3 \cup \{e_1, e_2\}$ with the obvious order. For a sequence E, the same set with the reverse order is denoted by E^R. The cost of the optimal alphabetic tree on E is denoted by $A(E)$.

* Received by the editors October 9, 1979, and in revised form October 22, 1980. This research was supported in part by the U.S. Office of Naval Research under contract N00014-76-C-0366.

† Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

‡ Current address, Department of Mathematics, University of California, Los Angeles, California 90024.

In an E-tree, a nonleaf is called an *interior node*. A node lying on the path between a given node n and the root is an *ancestor* of n and the first such node is the *father* of n. Every interior node is the father of two *sons*, referred to as the right and left sons if the tree is embedded in the plane (alphabetic trees are always assumed to have the embedding which gives the required leaf order). The *subtree rooted at* a node n is the tree with root n consisting of n and its descendants. For any node n, $L(n)$ is the set of leaves in the subtree rooted at n.

2. Some facts about optimal alphabetic trees. In this section we present some facts about alphabetic trees which are needed to prove the theorem. For a more complete discussion the reader is referred to [4].

Given the ordered set E, the Hu–Tucker algorithm constructs a rooted binary tree T with leaf set E which, though not necessarily alphabetic, satisfies $d_T(e) = d_{T'}(e)$ for all $e \in E$, where T' is the optimal alphabetic tree for E. Starting with the sequence E of leaves, the algorithm assigns a father to two nodes a and b, removes them from the sequence and inserts their father (a, b) in the sequence where the smaller weighted of a and b was. This node is assigned a weight of $w(a) + w(b)$. This operation, called a *merge* of a and b, is repeated until one node remains; this remaining node is the root and the tree is completed. At each stage, the choice of which pair of nodes to merge is made as follows. Two nodes in the sequence are said to be *compatible* if they are adjacent in the sequence or separated only by nonleaf nodes. The pair of nodes a and b which are merged is a *locally minimum compatible pair* (l.m.c.p. for short), which has the property that b has the smallest weight of those nodes compatible with a and a has the smallest weight of those nodes compatible with b. There is always at least one l.m.c.p., the smallest weight element together with the minimum weight element with which it is compatible. If there is more than one l.m.c.p., any one can be merged; it is a property of the algorithm that the resulting tree will have the desired properties for any sequence of l.m.c.p. merges.

The above construction can be used to prove several lemmas about the optimal alphabetic tree.

LEMMA 1. (i) *If* $(E_1, e_1, E_2, e_2, E_3)$ *is a sequence with the weights of* e_1 *and* e_2 *exceeding the weight of any element in* E_2, *then*

$$A(E_1, e_1, E_2, e_2, E_3) = A(E_1, e_1, E_2^R, e_2, E_3).$$

(ii) *If in the sequence* (E_1, e_1, E_2) *the weight of* e_1 *exceeds the weight of any element in* E_2, *then*

$$A(E_1, e_1, E_2) = A(E_1, e_1, E_2^R).$$

Proof. (i) We show that the Hu–Tucker algorithm performs the same set of merges in both sequences, so that each node is at the same depth in the optimal tree for both. If $|E_2| = 1$, the result is trivial. Otherwise, E_2 contains an l.m.c.p. which is also an l.m.c.p. in E_2^R, so it is merged in both. This process is repeated merging l.m.c.p.'s which lie between e_1 and e_2 until there is no l.m.c.p. between e_1 and e_2. This can happen only when there is at most one leaf remaining between e_1 and e_2 in the sequence. If there are no leaves remaining, the compatibilities occurring in both sequences are now the same. If there is one leaf e remaining, but no l.m.c.p., then e has smaller weight than any of the (nonleaf) nodes lying between e_1 and e_2. Perform all l.m.c.p. merges not involving the nodes between e_1 and e_2 (these are obviously the same in both sequences). After this, there is an l.m.c.p. involving at least one node lying between e_1 and e_2, which must be e since anything compatible to a node lying between e_1 and e_2 is compatible to e.

Moreover, e has the same compatibilities in both sequences, so this l.m.c.p. can be merged in both. Once e merges, the compatibilities in the two sequences are identical and the algorithm will act in a parallel manner on the two sequences.

The proof of (ii) is essentially the same. □

LEMMA 2. *Let (E, e, F, f, G) be a sequence of weighted elements such that all elements in F have smaller weight than e and f. Let s be the smallest weight element in F. Then there is an ordering (F', s) of F such that*

$$A(E, e, F', s, f, G) = A(E, e, F, f, G).$$

Proof. The proof is by induction on $|F|$. If $|F| = 1$, the result is trivial. Otherwise, let m be the element in F of largest weight and write F as (F_1, m, F_2), so that the given sequence is $(E, e, F_1, m, F_2, f, G)$. If $s \in F_2$ we can apply induction since $|F_2| < |F|$; if $s \in F_1$ we can apply Lemma 1 to reverse F and then use induction on F_1^R. □

LEMMA 3. *If (E_1, e_1, E_2) is a leaf sequence with the weight of e_1 greater than that of every element in E_2, then in the optimal alphabetic tree T, $d_T(e) \geqq d_T(e_1) - 1$ for all $e \in E_2$.*

Proof. Suppose the lemma does not hold, and let $f \in E_2$ be the first element of E_2 such that $h = d_T(f) < d_T(e_1) - 1$. Then a cheaper alphabetic tree can be constructed as follows. Let g be the ancestor of e_1 at height $h+1$ with n_0 and n_1 its left and right sons. Let $n_2, \cdots n_{k-1}$ be the sequence of nodes at height $h+1$ to the left of f and to the right of g and set $n_k = f$. Detach n_1 (and its subtree), reducing the height of the subtree rooted at n_0 by 1, and detach $n_2, n_3, \cdots n_k$ (and their subtrees) from T. Attach n_i to where n_{i+1} was for $1 \leqq i \leqq k-2$ and merge n_{k-1} and n_k, attaching them to where n_k was. In the resulting tree, nodes $n_2, \cdots, n_{k-1}$ remain at depth $h+1$ but the depth of f is increased by 1 and the depth of n_0 and n_1 are decreased from $h+2$ to $h+1$. Since e_1 is a leaf of n_0 or n_1 and $w(e_1) > w(f)$, the resulting tree is cheaper, contradicting the minimality of T. □

LEMMA 4. *If T is an alphabetic tree for (E, F), then there exists an alphabetic tree T' for F such that*

$$c(T') \leqq \sum_{e \in F} d_T(e) w(e) - \min_{e \in F} w(e).$$

Proof. The elements of E can be removed one at a time from T. Each time an element is removed, the other node with the same father is elevated one level, reducing the depth of all nodes in its subtree by 1. Removing the final node of E must elevate some element in F, reducing its depth by 1. Thus, the resulting tree T' has cost bounded by the above expression. □

LEMMA 5. *Let E_1 and E_2 be sequences and a an element with smaller weight than any in E_1 or E_2. Then*

$$A(E_1, a) + A(E_2) \leqq A(E_1, E_2).$$

Proof. Let T be an optimal alphabetic tree for (E_1, E_2). We will construct alphabetic trees T_1 for (E_1, a) and T_2 for E_2, such that $c(T) \geqq c(T_1) + C(T_2)$.

Let f be the final element of E_2 and let p be the ancestor of f in T which is also an ancestor of an element in E_1 and such that $d_T(p)$ is maximum. Let s_R be the right son of P; s_R is an ancestor of f and $d_T(s_R) > d_T(s)$, so, by choice of p, s_R precedes only elements of E_2. Construct T_1 by removing the subtree rooted at s_R from T and replacing it by a, then reducing the tree, according to Lemma 4, by removing any remaining elements of E_2. The depth in T_1 of every element of E_1 is no greater than in T

and a is at depth $d_T(p)+1$, so

$$c(T_1) \leqq \sum_{e \in E_1} d_T(e)w(e) + (d_T(p)+1)w(a). \tag{1}$$

Since T is alphabetic and p is an ancestor of both f and an element of E_1, it must be an ancestor of every element in E_2. Let T_2 be the tree obtained from the subtree T_p rooted at p by deleting all elements of E_1. By Lemma 4, $c(T_2) \leqq \sum_{e \in E_2} d_{T_p}(e)w(w) - \min_{e \in E_2} w(e)$. Since for all $e \in E_2$, $d_{T_p}(e) = d_T + d_T(p)$, we have

$$\begin{aligned} c(T_2) &\leqq \left[\sum_{e \in E_2} d_T(e)w(e)\right] - \left[d_T(p)\left(\sum_{e \in E_2} w(e)\right)\right] - \min_{e \in E_2} w(e) \\ &\leqq \left[\sum_{e \in E_2} d_T(e)w(e)\right] - \left[(d_T(p)+1)\left(\min_{e \in E_2} w(e)\right)\right]. \end{aligned} \tag{2}$$

Combining (1) and (2), we have

$$\begin{aligned} c(T_1)+c(T_2) &\leqq \left[\sum_{e \in E_1 \cup E_2} d_T(e)w(e)\right] + (d_T(p)+1)w(a) \\ &\quad - \left[(d_T(p)+1)\min_{e \in E_2} w(e)\right] \\ &\leqq c(T) + (d_T(p)+1)(w(a) - \min_{e \in E_2} w(e)) \\ &\leqq c(T), \end{aligned}$$

since $w(a) < w(e)$ for all $e \in E_2$. □

3. Main theorem.

THEOREM. *Let $(e_0, e_1, \cdots, e_t)$ be a set indexed from smallest to largest weight. Then the linear order of the set requiring the most expensive alphabetic tree is*

$$(e_0, e_t, e_1, e_{t-1}, \cdots, e_j, e_{t-j}, \cdots).$$

Proof. Let E be the most expensive sequencing of the elements and let i be the largest integer such that the first $2i$ elements of E are $e_0, e_t, e_1, e_{t-1}, \cdots, e_{i-1}, e_{t-i+1}$ (if the first two elements of E are not e_0, e_t, then $i = 0$). Call this subsequence B and the remaining subsequence F, so $E = (B, F)$. If $|F| \leqq 1$, E is as prescribed by the theorem. For $|F| > 1$, we show that there exists a reordering (e_i, e_{t-i}, F') of F such that $A(B, F) \leqq A(B, e_i, e_{t-i}, F')$ and the theorem will follow by induction. Write F as (F_1, e_{t-i}, F_2) and suppose $e_i \in F_1$ (if not, apply Lemma 1 (ii) to reverse F). By Lemma 2, there is a reordering (F_1', e_i) of F_1 such that $(B, F_1', e_i, e_{t-i}, F_2)$ is the same cost as E. Let T be the optimal tree for $(B, e_i, e_{t-i}, F_1', F_2)$; we construct an alphabetic tree for $(B, F_1', e_i, e_{t-i}, F_2)$ which is no more expensive than T, which will prove the theorem.

The pair (e_i, e_{t-i}) is an l.m.c.p. in $(B, e_i, e_{t-i}, F_1', F_2)$, so it can be merged first in the Hu–Tucker algorithm.Therefore, in the final Hu–Tucker tree (and, hence, also in T) e_i and e_{t-i} are at the same depth h. Lemma 3 implies that every leaf to the right of e_{t-i} is at depth at least $h-1$.

Case 1. e_i and e_{t-i} have the same father n_0 in T. Let $n_1, \cdots, n_k$ be the sequence of nodes at depth $h-1$ in T which lie to the right of n_0 and let n_c be the first of these such that $L(n_c) \cup F_2 \neq \varnothing$. In the tree, permute the subtrees rooted at $n_0, \cdots, n_{c-1}$ by attaching n_1 where n_0 was, n_2 where n_1 was, . . . , n_{c-1} where n_{c-2} was and n_0 where n_{c-1}

was. The resulting tree T^* has the same cost as T. If $L(n_c) \subseteq F_2$, then the order of the leaves in T^* is $(B, F_1', e_i, e_{t-i}, F_2)$ and we are done. Otherwise, let S_L and S_R be the left and right sons of n_c. For convenience, the positions occupied by e_i, e_{t-i}, S_L and S_R in T^* will be referred to, respectively, as P_1, P_2, P_3 and P_4.

There are three possibilities to consider:

(a) $L(S_L) \subseteq F_1'$, $L(S_R) \subseteq F_2$;

(b) $L(S_L) \subseteq F_1'$, $L(S_R)$ intersects both F_1' and F_2;

(c) $L(S_L)$ intersects both F_1' and F_2, $L(S_R) \subseteq F_2$.

For (a), attach S_L to P_1, e_i to P_2, e_{t-i} to P_3 and S_R to P_4. The result is alphabetic for $(B, F_1', e_i, e_{t-i}, F_2)$ and the same cost as T^*.

For (b) write $L(S_R)$ as (G_1, G_2) where $G_1 \subseteq F_1'$ and $G_2 \subseteq F_2$. By Lemma 5, there are alphabetic trees T_1 for (G_1, e_i) and T_2 for G_2 such that $c(T_1) + c(T_2)$ does not exceed the cost of the subtree rooted at S_R. Hence, attaching S_L to P_1, T_1 to P_2, e_{t-i} to P_3, and T_2 to P_4 yields an alphabetic tree for $(B, F_1', e_i, e_{t-i}, F_2)$ which is cheaper than T^*.

For (c), do for S_L what was done for S_R in (b) and assign T_1 to P_1, e_{t-i} to P_2, T_2 to P_3 and S_R to P_4.

Case 2. e_i and e_{t-i} have different fathers. Let S_R be the right son of the father of e_{t-i}. If $L(S_R) \subseteq F_1'$, then attach S_R to where e_i was, e_i to where e_{t-i} was and e_{t-i} to where S_R was, and now Case 1 applies. Otherwise, write $L(S_R) = (G_1, G_2)$, where $G_1 \subseteq F_1'$, $G_2 \subseteq F_2$, and construct trees T_1 and T_2 as in (ii) of Case 1. Attach T_1 to where e_i was and T_2 to where S_R was, and the result is cheaper than T and alphabetic for (B, F_1', e_i, F_2).

Acknowledgments. The authors would like to thank T. C. Hu for suggesting the problem and encouraging this work, and M. Wachs for pointing out some errors in the proofs of Lemmas 1 and 3 in an earlier version of this paper. We are also grateful to P. Edelman, G. W. Peck, J. Shearer and D. Sturtevant for stimulating conversations.

REFERENCES

[1] A. M. GARSIA AND M. L. WACHS, *A new algorithm for minimal binary search trees*, SIAM J. Comput., 6 (1977), pp. 622–642.

[2] E. N. GILBERT AND E. F. MORE, *Variable length binary encodings*, Bell System Tech. J., 38 (1959), pp. 933–968.

[3] T. C. HU, *A new proof of the T-C algorithm*, SIAM J. Appl. Math., 25 (1973), pp. 83–94.

[4] T. C. HU, D. J. KLEITMAN AND J. K. TAMAKI, *Binary trees optimum under various criteria*, SIAM J. Appl. Math., 37 (1979), pp. 246–256.

[5] T. C. HU AND A. C. TUCKER, *Optimal computer-search trees and variable-length alphabetical codes*, SIAM J. Appl. Math., 21 (1971), pp. 514–432.

[6] D. A. HUFFMAN, *A method for the construction of minimum redundancy codes*, Proc. IRE, 40 (1952), pp. 1098–1101.

[7] D. E. KNUTH, *Fundamental algorithms*, Vols. 1 and 3, Addison-Wesley, Reading, MA, 1973.

[8] J. H. VAN LINT, *Coding Theory*, No. 201, Springer-Verlag, Berlin, 1973.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 2, June 1981

0196-5212/81/0202-0009 $01.00/0

A TIGHT ASYMPTOTIC BOUND FOR NEXT-FIT-DECREASING BIN-PACKING*

B. S. BAKER† AND E. G. COFFMAN, JR.†

Abstract. In this note we derive a tight asymptotic bound on the relative performance of the Next-Fit-Decreasing approximation rule for classical one-dimensional bin-packing. The proof provides a novel application of certain well-known sequences of unit fractions. Potential applications are mentioned.

1. Introduction. Bin-packing problems model a large variety of practical problems arising in computer sciences and operations research. The general problem assumes a collection of equal capacity bins and a list of pieces which are to be packed into the bins subject to the requirement that the capacity of no bin be exceeded. (As usual, the pieces are assumed to have sizes not exceeding the common bin capacity.) The specific problem, which is NP-complete [3], is to minimize the number of bins used in the packing. A number of approximation algorithms have been analyzed for this problem with the objective of characterizing worst-case performance relative to optimal packings. This paper is devoted to a similar analysis of one such algorithm, called the Next-Fit-Decreasing (NFD) rule.

Broadly speaking, three basic bin-packing algorithms can be identified: Next-Fit (NF), First-Fit (FF) and Best-Fit (BF). To define these approximation rules let B_1, $B_2, \cdots$ be an arbitrary ordering of the bins, and let $L = (p_1, p_2, \cdots, p_n)$ denote the list of pieces to be packed, with the convention that they are to be packed in the order given. Without loss of generality we assume unit bin capacities so that $p_i \in (0, 1]$ for all i. The Next-Fit (NF) rule begins by placing as many of $p_1, p_2, \cdots$ into B_1 as can be done without exceeding the bin capacity. If we assume that $p_1, \cdots, p_i$, $i < n$, are thus packed into B_1, the next step is to pack as many of $p_{i+1}, \cdots, p_n$ as possible into B_2. This process is repeated with B_3 and so on until the last piece is packed.

The FF rule places each successive piece into the first (leftmost) bin of the sequence $B_1, B_2, \cdots$ into which it will fit. The BF rule places each successive piece into the leftmost bin for which the resulting unused capacity is the least; i.e., each piece goes into a smallest, sufficiently large hole (unused capacity) that can be found at the time it is packed; ties are resolved in favor of the bin of lowest index. Note that according to FF and BF it is generally possible for a piece to be packed to the left of the rightmost occupied bin. But this does not occur with NF, which fills the bins in sequence; i.e., $B_1, \cdots, B_{i-1}$ receive no further pieces after the first piece is packed in B_i.

Important variations of these rules are obtained by augmenting each with an initial arrangement of L into nonincreasing order of piece size. These variations are denoted NFD, FFD and BFD, with the D standing for "Decreasing".

The analysis of approximation rules has concentrated on the derivation of worst-case bounds of the form $A(L) \leqq \alpha \mathrm{OPT}(L) + \beta$, where α and β are constants and $A(L)$ and $\mathrm{Opt}(L)$ are the numbers of bins required to pack L by algorithm A and an optimization rule, respectively. The multiplicative constant is an asymptotic bound on $A(L)/\mathrm{OPT}(L)$, and it is the main focus of the analysis. For all but NFD of the six rules we have defined, tight asymptotic bounds have been known for several years [5], [6]; these are $\alpha = 2$ for NF, $\frac{17}{10}$ for FF and BF and $\frac{11}{9}$ for FFD and BFD. In the next section we

* Received by the editors July 14, 1980.
† Bell Laboratories, Murray Hill, New Jersey 07974.

show that the sum

$$\gamma = \sum_{i=1}^{\infty} \frac{1}{a_i} = 1 + \tfrac{1}{2} + \tfrac{1}{6} + \tfrac{1}{42} + \cdots \approx 1.691,$$

where $a_1 = 1$, $a_{i+1} = a_i(a_i + 1)$, $i \geqq 1$, is a tight asymptotic bound for NFD.

For specific illustrations of the wide variety of applications served by bin-packing models, the reader is referred to [6]. From the description given earlier, note that the particular applications of the simpler NF rule correspond to situations in which a strict, sequential processing of the bins is required, and ordering the input, L, is either impossible or too costly to implement. The more effective FFD and BFD rules will apply whenever L can be ordered by piece size and strict sequencing of bins is not required, or whenever the input can be put initially into that order from which a packing of the bins in sequence produces the corresponding FFD or BFD packing. The NFD packings, less effective than FFD and BFD, appear to be limited to those NF situations where ordering by size is possible, but not by precomputed FFD or BFD packings, a situation that might occur if the bin size is not known when the input can be ordered. In this connection, it is perhaps best to consider our NFD model as a special case of the more general problem where bin sizes vary unpredictably, so that packings cannot be determined in advance.

Apart from potential applications, the proof of the NFD results lends further significance to an interesting sequence, the a_i's defined above. This sequence has arisen in other research in bin-packing problems [2], and it is closely related to classes of sequences studied by Golomb [4] and Aho and Sloane [1].

2. The NFD bound. In the main result to follow, we provide a general bound in which maximum piece size is a parameter (similar results can be found in [5], [6] for the other rules). For this purpose we shall be using two simply related classes of sequences. First, for $s \geqq 1$ an integer, let

$$t_1(s) = s + 1, \qquad t_2(s) = s + 2,$$
$$t_{i+1}(s) = t_i(s)[t_i(s) - 1] + 1, \qquad i \geqq 2.$$

For example, the first few integers of the first three sequences are

$$2,\ 3,\ 7,\ 43,\ 1{,}807, \cdots$$
$$3,\ 4,\ 13,\ 157,\ 24{,}493, \cdots$$
$$4,\ 5,\ 21,\ 421,\ 17{,}6821, \cdots$$

Next, define the *γ_s-sequences*, $\{a_i(s)\}$,

$$a_i(s) = t_i(s) - 1, \qquad i \geqq 1,$$

and let $\gamma_s = \sum_{i=1}^{\infty} 1/a_i(s) = \sum_{i=1}^{\infty} 1/(t_i(s) - 1)$. Note that the sequence $\{a_i\}$ given earlier is the γ_1-sequence, and $\gamma = \gamma_1$. It is easily verified that $\sum_{i=1}^{\infty} 1/t_i(s) = 2/(s+1)$ and that the tails corresponding to partial sums are given by

$$\frac{2}{s+1} - \sum_{i=1}^{k-1} \frac{1}{t_i(s)} = \frac{1}{t_k(s) - 1}.$$

THEOREM. *For list $L = (p_1, \cdots, p_n)$, $p_1 \geqq p_2 \geqq \cdots \geqq p_n$, let s be the smallest integer such that $p_1 \in (1/(s+1), 1/s]$. Then*

$$\text{NFD}\,(L) \leqq \gamma_s^*\ \text{OPT}\,(L) + 3,$$

where $\gamma_s^* = (s-1)/s + \gamma_s$. *Note that the* γ_s^* *are monotone decreasing. Moreover, the multiplicative constant* γ_s^* *is the smallest possible.*

Proof. We begin with the following notation. If $k = t_i(s) - 1$ for some i, then the interval $(1/(k+1), 1/k]$ is called a γ_s-interval. Pieces whose sizes are in γ_s-intervals will be called γ_s-pieces.

Next, we define below a weighting function of piece size, $W_s(p)$, which we shall prove has the following two properties. If $W_s(L)$ is the cumulative weight of the pieces in L, then 1) the length of the NFD packing of L cannot exceed $W_s(L)+3$; and 2) $W_s(L)$ can not exceed γ_s^* times the length of an optimum packing. The bound follows immediately from $\mathrm{NFD}(L) - 3 \leqq W_s(L) \leqq \gamma_s^* \mathrm{OPT}(L)$.

Define the weighting function $W_s(p)$ as follows. For $p \in (1/(k+1), 1/k]$, $k \geqq s$,

$$W_s(p) = \frac{1}{k} \qquad \text{if } k = t_i(s) - 1 \text{ for some } i \geqq 1,$$

$$= \frac{k+1}{k} p \quad \text{otherwise.}$$

Note that $W_s(p)$ is a nondecreasing function of p that is strictly increasing except in γ_s-intervals, where it remains constant. Note also that $W_s(p)/p$ decreases monotonically in γ_s-intervals but is a constant in any other interval.

CLAIM 1. $W_s(L) \geqq \mathrm{NFD}(L) - 3$.

Proof. If an NFD bin, B, has k pieces in the interval $(1/(k+1), 1/k]$, then the total weight, $W_s(B)$, of pieces in B is at least $k(1/k) = 1$, whether or not $(1/(k+1), 1/k]$ is a γ_s-interval. Otherwise, if B is not the last bin, it must contain pieces from at least two different intervals; such bins will be called transition bins. We now verify that the cumulative weight of these transition bins is at least two less than the total number of such bins. We divide the transition bins into two categories.

Case i. B is a transition bin with at least one γ_s-piece.

Let the smallest piece of B be in $(1/(k+1), 1/k]$, $k \geqq s+1$. By definition of the NFD rule the unoccupied space in B is no more than $1/k$, and hence B is at least $(k-1)/k$ full. Therefore, $W_s(B)$ is at least $(k-1)/k \min W_s(p)/p$; and since $W_s(p)/p \geqq 1$ for all $p > 0$ we have $W_s(B) \geqq 1 - 1/k$, and a maximum shortfall (amount less than 1) of $1/k$. Now at most two successive transition bins can have γ_s-pieces from the same γ_s-interval. Thus, the cumulative shortfall of the transition bins with γ_s-pieces is no greater than twice the sum of the numbers $1/k$ over all $k \geqq s+1$ such that $(1/(k+1), 1/k]$ is a γ_s-interval. Specifically, since the γ_s^* are decreasing in s we have

$$2 \sum_{i=2}^{\infty} \frac{1}{t_i(s) - 1} = 2\left(\gamma_s - \frac{1}{s}\right) = 2(\gamma_s^* - 1) \leqq 2(\gamma_1^* - 1) < \frac{3}{2}$$

as a bound on the shortfall due to transition bins with at least one γ_s-piece.

Case ii. The transition bin B contains no γ_s-piece.

If the smallest piece in B is in the interval $(1/(k+1), 1/k]$; then B is at least $(k-1)/k$ occupied, as before. Since B has no γ_s-pieces, $W_s(p)/p \geqq (k+1)/k$, the lower bound corresponding to the smallest piece. Thus, $W_s(B) \geqq (k-1)/k \cdot (k+1)/k = 1 - 1/k^2$, and hence the shortfall is at most $1/k^2$. Now the smallest piece in a transition bin with no γ_s-pieces can be no larger than $1/(s+3)$. Hence, the cumulative shortfall of such transition bins is at most the constant

$$\sum_{k \geqq s+3} \frac{1}{k^2} \leqq \frac{\pi^2}{6} - \frac{49}{36} < \frac{1}{3}.$$

Finally, the cumulative shortfall of all transition bins is at most $\frac{3}{2}+\frac{1}{3}<2$. Adding one for the maximum weight of the last NFD bin we have $W_s(L)\geqq \mathrm{NFD}(L)-3$. □

CLAIM 2. *In any packing of L the weight of any bin is at most γ_s^*. Hence, $W_s(L)\leqq \gamma_s^* \mathrm{OPT}(L)$.*

Proof. Suppose first that a bin B has $s-1$ pieces in the (largest) γ_s-interval $(1/(s+1),\ 1/s]$, with a cumulative weight of $(s-1)/s$, and total length at least $(s-1)/(s+1)$. Let $q_1\geqq q_2\geqq \cdots \geqq q_m$ be the remaining pieces in B, packed into an interval of length at most $1-(s-1)/(s+1)=2/(s+1)$.

If for all i, q_i is in the interval $(1/t_i(s),\ 1/(t_i(s)-1)]$, then

$$W_s(B)=\frac{s-1}{s}+\sum_{i=1}^{m}\frac{1}{t_i(s)-1}<\gamma_s^*.$$

Thus, suppose r is the least i such that $q_i\notin(1/t_i(s),\ 1/(t_i(s)-1)]$ and hence $q_r\leqq 1/t_r(s)$. The total weight of the largest $r-1$ pieces is $\sum_{i=1}^{r-1}1/(t_i(s)-1)$, and their total length is at least $\sum_{i=1}^{r-1}1/t_i(s)$. The remaining capacity is thus no more than $2/(s+1)-\sum_{i=1}^{r-1}1/t_i(s)=1/(t_r(s)-1)$. Since $q_r\leqq 1/t_r(s)$, we have $W_s(q_i)/q_i(s)\leqq (t_r(s)+1)/t_r(s)$ for all $i\geqq r$. Thus, the cumulative weight, W_s', of the remaining pieces $q_r, q_{r+1}, \cdots, q_m$ must satisfy

$$W_s'\leqq\frac{t_r(s)+1}{t_r(s)}\cdot\frac{1}{t_r(s)-1}=\frac{1}{t_r(s)-1}+\frac{1}{t_r(s)(t_r(s)-1)},$$

or

$$W_s'\leqq\frac{1}{t_r(s)-1}+\frac{1}{t_{r+1}(s)-1}.$$

Finally, therefore,

$$\begin{aligned}W_s(B)&\leqq\frac{s-1}{s}+\sum_{i=1}^{r-1}\frac{1}{t_i(s)-1}+W_s'\\&\leqq\frac{s-1}{s}+\sum_{i=1}^{r+1}\frac{1}{t_i(s)-1}<\gamma_s^*.\end{aligned}$$

It remains to consider the case when there are only $h\leqq s-2$ pieces with sizes in $(1/(s+1),\ 1/s]$. These pieces occupy a length of at least $h/(s+1)$ and have a total weight of h/s. The remaining pieces $\{q_i\}$ are packed in an interval of length at most $1-h/(s+1)$, and their sizes are at most $1/(s+1)$. Thus, $(W_s(q_i)/q_i)\leqq(s+2)/(s+1)$ for all i, and hence

$$\sum_{i=1}^{m}W_s(q_i)\leqq\max_{1\leqq i\leqq m}\frac{W_s(q_i)}{q_i}\sum_{i=1}^{m}q_i\leqq\frac{s+2}{s+1}\left(1-\frac{h}{s+1}\right)$$

and

$$W_s(B)\leqq\frac{h}{s}+\frac{s+2}{s+1}\left(1-\frac{h}{s+1}\right).$$

Since this is an increasing function of $h\leqq s-2$, we have

$$W_s(B)\leqq\frac{s-2}{s}+\frac{s+2}{s+1}\left(1-\frac{s-2}{s+1}\right)=\frac{s-2}{s}+\frac{3(s+2)}{(s+1)^2}.$$

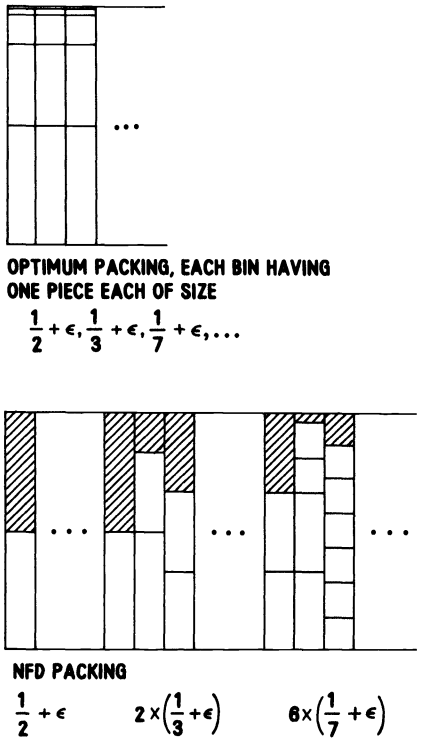


FIG. 1

But it is routine to verify that for all $s \geqq 1$

$$\frac{s-2}{s}+\frac{3(s+2)}{(s+1)^2} \leqq 1+\frac{1}{s+1}+\frac{1}{(s+1)(s+2)} < \gamma_s^*. \qquad \square$$

We complete the proof of the theorem by defining lists which show that the bound can be approached as closely as desired. To do this we need only formalize the construction maximizing bin weight as implied in Claim 2. Specifically, for $s \geqq 1$ and $k \geqq 1$ given, we construct the following optimum packing in $m = \mathrm{OPT}\,(L)$ bins. For an $\varepsilon > 0$ suitably small, each bin begins with $s-1$ pieces of size $1/(s+1)+\varepsilon$ and terminates with the sequence $1/t_1(s)+\varepsilon$, $1/t_2(s)+\varepsilon, \cdots, 1/t_k(s)+\varepsilon$ (note that $1/t_1(s)+\varepsilon = 1/(s+1)+\varepsilon$, so that we in fact have s such pieces). The packing is clearly a valid one, since the sum of piece sizes is

$$\frac{s-1}{s+1}+\sum_{i=1}^{k}\frac{1}{t_i(s)}+O(\varepsilon)<1$$

for a suitably small ε.

Now take this set of $m(k+s-1)$ pieces, put it into nonincreasing order, and apply NFD to the resulting list, L_k. Assuming that m is a sufficiently large multiple of t_k-1, the NFD packing begins with $m-1$ bins having s pieces of size $1/(s+1)+\varepsilon$ only, one transition bin, $m/(t_2(s)-1)-1$ bins having pieces of size $1/t_2(s)+\varepsilon$ only, another transition bin, etc., with the last bin having pieces of size $1/t_k(s)+\varepsilon$ only. An example is shown in Fig. 1 for $s=1$. Thus,

$$\mathrm{NFD}\,(L_k) \geqq m+m\sum_{i=2}^{k}\frac{1}{t_i(s)-1}-(k+1),$$

and the ratio

$$\frac{\mathrm{NFD}(L_k)}{\mathrm{OPT}(L_k)} \geqq 1 + \sum_{i=2}^{k} \frac{1}{t_i(s)-1} - \frac{k+1}{m}$$

$$= \frac{s-1}{s} + \sum_{i=1}^{k} \frac{1}{t_i(s)-1} - \frac{k+1}{m}$$

can be made as close to γ_s^* as desired by appropriate choices for k, m and ε. □

We note that the closeness with which we want to approach γ_s^* determines the minimum size of the list L_k. But because of the fast convergence of the series $\{1/t_i(s)\}$, we need only small values of k to achieve γ_s^* to within several decimal places. For example, if $s = 1$, then for $k \geqq 5$

$$\frac{\mathrm{NFD}(L_k)}{\mathrm{OPT}(L_k)} > 1.691.$$

The first few values of γ_s^* are given approximately by $\gamma_1^* = 1.691\cdots$, $\gamma_2^* = 1.423\cdots$, $\gamma_3^* = 1.302\cdots$, $\gamma_4^* = 1.233\cdots$, etc. Using the first two terms of the series, we see that γ_s^* approaches 1 with increasing a roughly as $(s+2)/(s+1)$.

REFERENCES

[1] A. V. AHO AND N. J. A. SLOANE, *Some doubly exponential sequences*, Fibonacci Quart., 11 (1973), pp. 429–437.

[2] 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.

[3] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability; A Guide to the Theory of* NP-*Completeness*, W. H. Freeman, San Francisco, 1979.

[4] S. GOLOMB, *On certain non-linear recurring sequences*, American Math. Monthly, 7 (1963), pp. 403–405.

[5] D. S. JOHNSON, *Fast algorithms for bin-packing*, J. Comput. System Sci., 8 (1974), pp. 272–314.

[6] 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. 256–278.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 2, June 1981

0196-5212/81/0202-0010 $01.00/0

RANDOM FLIGHTS ON REGULAR POLYTOPES*

LAJOS TAKÁCS†

Abstract. In a series of random flights a traveler visits the vertices of a regular polytope. The traveler starts at a given vertex and in each flight, independently of the others, chooses a vertex at random as the destination. In each flight the transition probability depends only on the distance between the starting vertex and the end vertex. In this paper we determine the probability that the traveler returns to the initial position at the end of the nth flight, and give explicit formulas for the n-step transition probabilities of a Markov chain describing the random flights.

1. Introduction. Let $\mathfrak{P}$ be a regular polytope with σ vertices. Denote by $\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_{\sigma-1}$ the rectangular Cartesian coordinates of the vertices. We assume that in a series of random flights a traveler visits the vertices of the polytope. The traveler starts at a given vertex and in each flight, independently of the others, chooses a vertex at random as the destination. In each flight the transition probability depends only on the distance between the starting vertex and the end vertex. Denote by $\mathbf{v}_n$ $(n = 1, 2, \cdots)$ the position of the traveler at the end of the nth flight and by $\mathbf{v}_0$ the initial position.

In the theory of probability it is an important problem to study the recurrence properties of various random flights such as the one described above. Most of these recurrence properties are determined by $p(n)$, the probability that the traveler returns to the initial position at the end of the nth flight. Here we are concerned with the determination of $p(n)$ for $n \geqq 0$. By symmetry we can choose any vertex, say $\mathbf{x}_0$, as the initial position. Then the problem is to determine

$$p(n) = \mathbf{P}\{\mathbf{v}_n = \mathbf{x}_0 | \mathbf{v}_0 = \mathbf{x}_0\} \tag{1}$$

for $n \geqq 0$.

The sequence $\{\mathbf{v}_n;\ n = 0, 1, 2, \cdots\}$ is a homogeneous Markov chain and we can determine $p(n)$ by calculating the n-step transition probabilities. If σ, the number of vertices, is large, it is not easy to determine the nth power of the transition probability matrix whose elements depend on several parameters. Fortunately, as we shall see, we can solve the problem in a simpler way too.

The results of this paper have their origin in a problem posed by G. Letac [6]. For a solution of this problem see L. Takács [14]. In a series of three papers G. Letac and L. Takács [7], [8], [9] studied random walks on various regular polytopes. In this paper, it will be demonstrated that for any polytope other than the four-dimensional 120-cell, the problem of finding $p(n)$ can be reduced to the particular case where the traveler always moves to an adjacent vertex. In addition, several surprising formulas will be derived for regular polytopes. These formulas reflect some interesting properties of the symmetry groups of the regular polytopes.

2. A Markov chain describing the random flights. Let us choose a fixed vertex, say $\mathbf{x}_0$, and divide the σ vertices of $\mathfrak{P}$ into disjoint sets in such a way that all vertices whose distances from $\mathbf{x}_0$ are the same belong to the same set. Denote by $d_1, d_2, \cdots, d_m$ the possible distances arranged in increasing order and let $d_0 = 0$. Define

$$S_j = \{\mathbf{x}_r : \|\mathbf{x}_r - \mathbf{x}_0\| = d_j\} \tag{2}$$

* Received by the editors July 9, 1979, and in revised form December 3, 1980.
† Department of Mathematics and Statistics, Case Western Reserve University, Cleveland, Ohio 44106.

for $j = 0, 1, \cdots, m$. The sets $S_0, S_1, \cdots, S_m$ are called the sections of $\mathfrak{P}$. Denote by

$$\sigma_j = N(S_j) \tag{3}$$

the number of vertices in the set S_j. Then

$$\sum_{j=0}^{m} \sigma_j = \sigma. \tag{4}$$

We assume that

$$\mathbf{P}\{\mathbf{v}_n = \mathbf{x}_s | \mathbf{v}_{n-1} = \mathbf{x}_r\} = p_j \tag{5}$$

if $\|\mathbf{x}_s - \mathbf{x}_r\| = d_j$ and $n = 1, 2, \cdots$, where $p_0, p_1, \cdots, p_m$ are given nonnegative numbers satisfying the requirement

$$\sum_{j=0}^{m} \sigma_j p_j = 1. \tag{6}$$

Let us define a sequence of random variables $\xi_0, \xi_1, \cdots, \xi_n, \cdots$ in such a way that $\xi_n = j$ if and only if $\mathbf{v}_n \in S_j$. We can demonstrate that for each regular polytope other than the four-dimensional 120-cell the sequence of random variables $\{\xi_n;\ n = 0, 1, 2, \cdots\}$ forms a homogeneous Markov chain with state space $I = \{0, 1, \cdots, m\}$ and transition probabilities

$$p_{ij} = \sum_{\nu=0}^{m} a_{ij\nu} p_\nu, \tag{7}$$

where $a_{ij\nu}$ is the number of subscripts $s = 0, 1, \cdots, \sigma - 1$ for which $\|\mathbf{x}_s - \mathbf{x}_0\| = d_j$ and $\|\mathbf{x}_s - \mathbf{x}_r\| = d_\nu$, and $\mathbf{x}_r$ is any vertex for which $\|\mathbf{x}_r - \mathbf{x}_0\| = d_i$. Briefly,

$$a_{ij\nu} = N\{s: s = 0, 1, \cdots, \sigma - 1, \|\mathbf{x}_s - \mathbf{x}_0\| = d_j, \|\mathbf{x}_s - \mathbf{x}_r\| = d_\nu, \|\mathbf{x}_r - \mathbf{x}_0\| = d_i\}. \tag{8}$$

The transition probability matrix

$$\boldsymbol{\pi} = [p_{ij}]_{i,j \in I} \tag{9}$$

can be expressed in the form

$$\boldsymbol{\pi} = \sum_{\nu=0}^{m} \mathbf{A}_\nu p_\nu, \tag{10}$$

where

$$\mathbf{A}_\nu = [a_{ij\nu}]_{i,j \in I}. \tag{11}$$

Obviously, $\mathbf{A}_0$ is the identity matrix.

If we arrange the n-step transition probabilities $p_{ij}^{(n)} (i \in I, j \in I)$ in the form of a matrix, then we get

$$[p_{ij}^{(n)}]_{i,j \in I} = \boldsymbol{\pi}^n, \tag{12}$$

and

$$p_{00}^{(n)} = p(n) \tag{13}$$

yields the desired probability (1).

In most of the cases m is much smaller than σ and thus it is easier to handle the Markov chain $\{\xi_n;\ n \geqq 0\}$ than $\{\mathbf{v}_n;\ n \geqq 0\}$. However, the transition probability matrix (9) still depends on the parameters $p_0, p_1, \cdots, p_m$, and at first sight it seems hopeless to

determine π^n for all $n \geqq 0$. Fortunately, several favorable circumstances make it possible to determine the Jordan decomposition of π in each case. Thus $p_{ik}^{(n)}$ can be determined explicitly.

The aforementioned method can also be applied to the four-dimensional 120-cell, but we should change somewhat the definition of the sections $S_0, S_1, \cdots, S_m$. Some of the sets S_j defined by (2) should be split into two or three disjoint subsets. Otherwise the procedure is similar.

In the following discussion we exclude the case of the four-dimensional 120-cells. This will be considered separately. For all other regular polytopes we determine explicitly the n-step transition probabilities $p_{ik}^{(n)}$ for $i \in I$ and $k \in I$.

3. Determination of the higher transition probabilities. In what follows we exclude the four-dimensional 120-cell, and $\mathfrak{P}$ denotes any other regular polytope. Denote by $\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_{\sigma-1}$ the rectangular Cartesian coordinates of the vertices of $\mathfrak{P}$ in a Euclidean space.

If we choose the center of the sphere which contains the vertices of $\mathfrak{P}$ as the origin of the coordinate system, then $\|\mathbf{x}_r\| = \rho$ for $r = 0, 1, \cdots, \sigma - 1$, where ρ is the circumradius of $\mathfrak{P}$, and for any two vertices of $\mathfrak{P}$ we have

$$\|\mathbf{x}_s - \mathbf{x}_r\|^2 = \|\mathbf{x}_s\|^2 + \|\mathbf{x}_r\|^2 - 2(\mathbf{x}_s, \mathbf{x}_r) = 2\rho^2 - 2(\mathbf{x}_s, \mathbf{x}_r), \tag{14}$$

where $(\mathbf{x}_s, \mathbf{x}_r)$ is the inner porduct of $\mathbf{x}_s$ and $\mathbf{x}_r$. Now we can replace (8) by the following equivalent definition:

$$a_{ij\nu} = N\{s: s = 0, 1, \cdots, \sigma - 1, (\mathbf{x}_s, \mathbf{x}_0) = c_j, (\mathbf{x}_s, \mathbf{x}_r) = c_\nu, (\mathbf{x}_r, \mathbf{x}_0) = c_i\}, \tag{15}$$

where

$$c_j = \rho^2 - \tfrac{1}{2}d_j^2 \tag{16}$$

and $\mathbf{x}_r$ is any vertex for which $(\mathbf{x}_r, \mathbf{x}_0) = c_i$. By choosing a suitable vertex $\mathbf{x}_r$, we can easily enumerate $a_{ij\nu}$ if we use (15).

From the definition of $a_{ij\nu}$ it follows that

$$\sigma_i a_{ij\nu} = \sigma_j a_{ji\nu} \tag{17}$$

and

$$a_{ij\nu} = a_{i\nu j} \tag{18}$$

for $i, j, \nu \in I$. These properties imply that $\sigma_i a_{ij\nu}$ is invariant under the six permutations of (i, j, ν).

Obviously, we have

$$\sum_{j \in I} a_{ij\nu} = \sigma_\nu \tag{19}$$

for $i, \nu \in I$. By (17) and (19) we get

$$\sum_{i \in I} \sigma_i a_{ij\nu} = \sigma_j \sum_{i \in I} a_{ji\nu} = \sigma_j \sigma_\nu \tag{20}$$

for $j, \nu \in I$.

If the polytope has central symmetry, then

$$\sigma_{m-\nu} = \sigma_\nu \tag{21}$$

for $\nu \in I$, and

$$a_{ij\nu} = a_{m-i,m-j,\nu} = a_{i,m-j,m-\nu} = a_{m-i,j,m-\nu} \tag{22}$$

for $i, j, \nu \in I$. We can easily derive these equations if we choose the vertices of $\mathfrak{P}$ so that $\mathbf{x}_{\sigma-r-1} = -\mathbf{x}_r$ for $r = 0, 1, \cdots, \sigma - 1$. The exceptional polytopes which have no central symmetry are the polygons with an odd number of sides and the simplexes in r dimensions ($r \geqq 2$).

Our aim is to determine the n-step transition probabilities $p_{ik}^{(n)}$ for the Markov chain $\{\xi_n; n = 0, 1, 2, \cdots\}$.

In what follows we shall use the notation $\mathbf{I}$ for the $(m+1) \times (m+1)$ identity matrix; that is,

$$\mathbf{I} = [\delta_{ij}]_{i,j \in I}, \tag{23}$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{24}$$

The eigenvalues of the matrix $\mathbf{A}_\nu$ ($\nu \in I$) arranged in some specific order will be denoted by $\lambda_{j\nu}$ ($j \in I$), and we define

$$\mathbf{\Lambda}_\nu = [\delta_{ij}\lambda_{j\nu}]_{i,j \in I} \tag{25}$$

for $\nu \in I$.

LEMMA 1. *The eigenvalues $\lambda_{j\nu}$ ($j \in I$) of $\mathbf{A}_\nu$ are real and satisfy the inequality $|\lambda_{j\nu}| \leqq \sigma_\nu$ for $j \in I$ and $\nu \in I$. There exists a real nonsingular matrix $\mathbf{H}_\nu$ such that*

$$\mathbf{A}_\nu \mathbf{H}_\nu = \mathbf{H}_\nu \mathbf{\Lambda}_\nu \tag{26}$$

for $\nu \in I$, where $\mathbf{\Lambda}_\nu$ is defined by (25).

Proof. Let

$$\mathbf{D} = [\delta_{ij}\sigma_j^{1/2}]_{i,j \in I}, \tag{27}$$

where the square root is positive. By (17) we have

$$\mathbf{D}^2 \mathbf{A}_\nu = \mathbf{A}_\nu' \mathbf{D}^2, \tag{28}$$

where $\mathbf{A}_\nu'$ is the transpose of $\mathbf{A}_\nu$. This implies that

$$\mathbf{D}\mathbf{A}_\nu \mathbf{D}^{-1} = \mathbf{D}^{-1}\mathbf{A}_\nu' \mathbf{D} = (\mathbf{D}\mathbf{A}_\nu \mathbf{D}^{-1})'. \tag{29}$$

Accordingly, the matrix

$$\mathbf{Q}_\nu = \mathbf{D}\mathbf{A}_\nu \mathbf{D}^{-1} \tag{30}$$

is symmetric. Obviously, $\mathbf{A}_\nu$ and $\mathbf{Q}_\nu$ have the same eigenvalues. The eigenvalues of $\mathbf{Q}_\nu$ are real and there exists an orthogonal matrix $\mathbf{M}_\nu$ such that

$$\mathbf{Q}_\nu \mathbf{M}_\nu = \mathbf{M}_\nu \mathbf{\Lambda}_\nu, \tag{31}$$

where $\mathbf{\Lambda}_\nu$ is defined by (25). In (31)

$$\mathbf{M}_\nu' \mathbf{M}_\nu = [\delta_{ij} m_j]_{i,j \in I}, \tag{32}$$

where m_j ($j \in I$) are arbitrary positive real numbers which we shall choose later in a convenient way.

Now the matrix

$$\mathbf{H}_\nu = \mathbf{D}^{-1}\mathbf{M}_\nu \tag{33}$$

is nonsingular, and by (30) and (31) it satisfies (26).

The elements of $\mathbf{A}_\nu$ are nonnegative and, by (19), in $\mathbf{A}_\nu$ each row-sum is σ_ν. Thus σ_ν is an eigenvalue of $\mathbf{A}_\nu$, and every eigenvalue of $\mathbf{A}_\nu$ has absolute value $\leqq\sigma_\nu$. This completes the proof of Lemma 1.

By (32) and (33) the inverse of $\mathbf{H}_\nu$ can be expressed by the transpose of $\mathbf{M}_\nu$ in the following form:

$$\mathbf{H}_\nu^{-1} = \mathbf{M}_\nu^{-1}\mathbf{D} = [m_j^{-1}\delta_{ij}]\mathbf{M}_\nu'\mathbf{D}. \tag{34}$$

The following observation is crucial in determining $\boldsymbol{\pi}^n$.

OBSERVATION 1. *The matrices* $\mathbf{A}_\nu$ $(\nu \in I)$ *commute in pairs*; *that is,*

$$\mathbf{A}_\mu\mathbf{A}_\nu = \mathbf{A}_\nu\mathbf{A}_\mu \tag{35}$$

if $\mu \in I$ *and* $\nu \in I$.

If we take into consideration that $\mathbf{Q}_\nu' = \mathbf{Q}_\nu$ where $\mathbf{Q}_\nu$ is defined by (30), then we can easily see that (35) holds if and only if

$$\mathbf{Q}_\mu\mathbf{Q}_\nu = \mathbf{Q}_\nu\mathbf{Q}_\mu, \tag{36}$$

or if and only if

$$\mathbf{Q}_\mu\mathbf{Q}_\nu = \mathbf{D}\mathbf{A}_\mu\mathbf{A}_\nu\mathbf{D}^{-1} \tag{37}$$

is a symmetric matrix.

In checking (35) it is convenient to use the second criterion. For the matrix (37) is symmetric if and only if $\sigma_i c_{ij} = \sigma_j c_{ji}$ for $i \in I$ and $j \in I$ where c_{ij} is the (i, j)-entry of $\mathbf{A}_\mu\mathbf{A}_\nu$.

Observation 1 implies that there exists an orthogonal matrix $\mathbf{M}$ such that, if $\mathbf{M}_\nu = \mathbf{M}$ for $\nu \in I$, then (31) and (32) are satisfied for all $\nu \in I$.

If we define

$$\mathbf{H} = \mathbf{D}^{-1}\mathbf{M}, \tag{38}$$

then by (26)

$$\mathbf{A}_\nu\mathbf{H} = \mathbf{H}\boldsymbol{\Lambda}_\nu \tag{39}$$

for each $\nu \in I$, and

$$\mathbf{M}'\mathbf{M} = [\delta_{ij}m_j] \tag{40}$$

is a diagonal matrix whose diagonal elements are positive real numbers.

By (38) and (40) we have

$$\mathbf{H}^{-1} = [m_j^{-1}\delta_{ij}]\mathbf{H}'\mathbf{D}^2. \tag{41}$$

If we know $\mathbf{H}$, then the eigenvalues of $\mathbf{A}_\nu$ $(\nu \in I)$ can be determined simply by matrix multiplication. For by (39) we have

$$\boldsymbol{\Lambda}_\nu = \mathbf{H}^{-1}\mathbf{A}_\nu\mathbf{H} \tag{42}$$

if $\nu \in I$ and $\mathbf{H}^{-1}$ can be calculated (41). By (42) the eigenvalues of $\mathbf{A}_\nu$, $\lambda_{j\nu}$ $(j \in I)$, are automatically arranged for $\nu \in I$ in some matching order.

Observation 1 implies that the eigenvalues of $\boldsymbol{\pi}$ are linear combinations of $p_0, p_1, \cdots, p_m$. For by (10) and (39) we have

$$\boldsymbol{\pi}\mathbf{H} = \mathbf{H}\mathbf{L}, \tag{43}$$

where

$$\mathbf{L} = \sum_{\nu=0}^{m} p_\nu\boldsymbol{\Lambda}_\nu = [\delta_{ij}\lambda_j]_{i,j\in I} \tag{44}$$

is a diagonal matrix with diagonal elements

$$\lambda_j = \sum_{\nu=0}^{m} p_\nu \lambda_{j\nu} \tag{45}$$

for $j \in I$. Hence we can conclude that λ_j $(j \in I)$ are the eigenvalues of $\boldsymbol{\pi}$, and

$$\boldsymbol{\pi} = \mathbf{H}\mathbf{L}\mathbf{H}^{-1} \tag{46}$$

is a Jordan decomposition of $\boldsymbol{\pi}$. The n-step transition probabilities, $p_{ik}^{(n)}$, are the elements of the matrix

$$\boldsymbol{\pi}^n = \mathbf{H}\mathbf{L}^n\mathbf{H}^{-1}. \tag{47}$$

Accordingly, we reduced the problem of finding $p_{ik}^{(n)}$ to the problem of finding the matrix $\mathbf{H}$ and the eigenvalues of $\mathbf{A}_\nu$ for $\nu \in I$. It is worthwhile to point out that, while the problem of finding the eigenvalues of $\boldsymbol{\pi}$ is an algebraic problem (the elements of $\boldsymbol{\pi}$ depend on the parameters $p_0, p_1, \cdots, p_m$), the problem of finding the eigenvalues of $\mathbf{A}_\nu$ $(\nu \in I)$ is a numerical problem (the elements of $\mathbf{A}_\nu$ are given nonnegative integers).

The eigenvalues $\lambda_{j\nu}$ $(j \in I, \nu \in I)$ are completely determined by the matrix

$$\mathbf{H} = [h_{ij}]_{i,j \in I} \tag{48}$$

defined by (38). If we form the $(0, j)$-entry on both sides of (39), we get

$$\sigma_\nu h_{\nu j} = h_{0j}\lambda_{j\nu}. \tag{49}$$

Since $\sigma_\nu > 0$ for all $\nu \in I$ and since $h_{\nu j} \neq 0$ for some $\nu \in I$, therefore $h_{0j} \neq 0$. Thus we obtain that

$$\lambda_{j\nu} = \sigma_\nu \frac{h_{\nu j}}{h_{0j}} \tag{50}$$

for $j \in I$, $\nu \in I$. This is a very simple formula for the determination of the eigenvalues of $\mathbf{A}_\nu$ $(\nu \in I)$.

It remains to solve the problem of how to choose $\mathbf{H}$ such that (39) will hold for all $\nu \in I$. Fortunately, the next observation provides an easy solution.

OBSERVATION 2. *The eigenvalues of* $\mathbf{A}_1$ *are distinct.*

This implies that in the equation

$$\mathbf{A}_1\mathbf{H} = \mathbf{H}\boldsymbol{\Lambda}_1 \tag{51}$$

the elements of the matrix (48) are determined up to a nonzero factor depending only on j. Consequently, $h_{\nu j}/h_{0j}$ is uniquely determined by $\mathbf{A}_1$. We can choose $h_{0j} \neq 0$ as we please.

Thus we have demonstrated that the problem of finding the Jordan decomposition of $\boldsymbol{\pi}$ can be reduced to the problem of finding the Jordan decomposition of $\mathbf{A}_1$. Stated more simply, if we know the matrix $\mathbf{H}$ for a regular polytope, then we have all the information needed to determine the n-step transition probabilities $p_{ik}^{(n)}$.

We can deduce the statement in Observation 1 from some general properties of the symmetry groups of the regular polytopes considered, and we can also extend the results of this paper to random walks on groups which share the characteristic properties of the symmetry groups of the regular polytopes.

In what follows we assume that $\mathbf{H}$ is a nonsingular matrix satisfying (39), and the elements of

$$\mathbf{\Lambda} = [\lambda_{j\nu}]_{j,\nu\in I} \tag{52}$$

are the eigenvalues of $\mathbf{A}_\nu$ for $\nu \in I$.

First, we express the quantities $a_{ik\nu}$ with the aid of the elements of the matrix $\mathbf{H}$.

THEOREM 1. *We have*

$$a_{ik\nu} = \sigma_k \sigma_\nu \sum_{j\in I} \omega_j \left(\frac{h_{ij}}{h_{0j}}\right)\left(\frac{h_{kj}}{h_{0j}}\right)\left(\frac{h_{\nu j}}{h_{0j}}\right) \tag{53}$$

for $i \in I$, $k \in I$, $\nu \in I$, where

$$\omega_j = \left[\sum_{i\in I} \sigma_i \left(\frac{h_{ij}}{h_{0j}}\right)^2\right]^{-1}. \tag{54}$$

Proof. By (50),

$$\mathbf{\Lambda} = \left[\sigma_\nu \frac{h_{\nu j}}{h_{0j}}\right]_{j,\nu\in I} = \left[\frac{\delta_{ij}}{h_{0j}}\right]\mathbf{H}'\mathbf{D}^2, \tag{55}$$

and, by (38),

$$\mathbf{H}' = \mathbf{M}'\mathbf{D}^{-1}. \tag{56}$$

Thus

$$\mathbf{\Lambda H} = \left[\frac{\delta_{ij}}{h_{0j}}\right]\mathbf{M}'\mathbf{M} = \left[\delta_{ij}\frac{m_j}{h_{0j}}\right]_{i,j\in I} \tag{57}$$

is a diagonal matrix. On the other hand, we have

$$\mathbf{\Lambda H} = \left[\sigma_j \frac{h_{ji}}{h_{0i}}\right]_{i,j\in I} [h_{jk}]_{j,k\in I}. \tag{58}$$

By comparing the (k, k)-entries in the matrices (57) and (58) we get the identity

$$\sum_{j\in I} \sigma_j \left(\frac{h_{jk}}{h_{0k}}\right)^2 = \frac{m_k}{(h_{0k})^2} \tag{59}$$

for $k \in I$. The left-hand side of (59) is independent of the particular choice of $\mathbf{H}$. If we choose $h_{0k} \neq 0$ as we please, then by (59) we get

$$m_k = \frac{(h_{0k})^2}{\omega_k}, \tag{60}$$

where ω_k is defined by (54). This answers the question of how to choose m_k in (40) or in (32).

By (41) and (60) we have

$$\mathbf{H}^{-1} = \left[\delta_{ij}\frac{\omega_j}{(h_{0j})^2}\right]_{i,j\in I}\mathbf{H}'\mathbf{D}^2. \tag{61}$$

We note that by (57)

$$\mathbf{\Lambda H} = \left[\delta_{ij}\frac{h_{0j}}{\omega_j}\right]_{i,j\in I}, \tag{62}$$

and consequently

$$\mathbf{\Lambda}^{-1} = \mathbf{H}\left[\delta_{ij}\frac{\omega_j}{h_{0j}}\right]_{i,j\in I}. \tag{63}$$

By (39), (48), (50) and (61) we obtain that

$$\begin{aligned}\mathbf{A}_\nu = \mathbf{H}\mathbf{\Lambda}_\nu\mathbf{H}^{-1} &= [h_{ij}]\left[\delta_{ij}\sigma_\nu\frac{h_{\nu j}}{h_{0j}}\right]\left[\delta_{ij}\frac{\omega_j}{(h_{0j})^2}\right][h_{ji}][\delta_{ij}\sigma_j] \\ &= \sigma_\nu[h_{ij}]\left[\delta_{ij}\frac{\omega_j h_{\nu j}}{(h_{0j})^3}\right][h_{ji}][\delta_{ij}\sigma_j].\end{aligned} \tag{64}$$

The (i, k)-entry of (64) is given by (53). This completes the proof of Theorem 1.

By Theorem 1 we can conclude that if the numbers a_{ik1} are known for $i \in I$, $k \in I$, then the numbers $a_{ik\nu}$ are uniquely determined for all $i \in I$, $k \in I$ and $\nu \in I$. It would be interesting to express $a_{ik\nu}$ directly as a function of a_{ik1} $(i \in I, k \in I)$ without the intervention of the matrix $\mathbf{H}$.

Finally, we can express the n-step transition probabilities in an explicit form with the aid of the matrix $\mathbf{H}$.

THEOREM 2. *The n-step transition probabilities of the Markov chain* $\{\xi_n;\ n = 0, 1, 2, \cdots\}$ *are given by the following formula*:

$$p_{ik}^{(n)} = \sigma_k \sum_{j\in I} \omega_j\left(\frac{h_{ij}}{h_{0j}}\right)\left(\frac{h_{kj}}{h_{0j}}\right)\lambda_j^n, \tag{65}$$

where

$$\lambda_j = \sum_{\nu\in I} \lambda_{j\nu}p_\nu = \sum_{\nu\in I} p_\nu\sigma_\nu\frac{h_{\nu j}}{h_{0j}} \tag{66}$$

and ω_j *is defined by* (54).

Proof. By (47) and (61) we have

$$\boldsymbol{\pi}^n = \mathbf{H}[\delta_{ij}\lambda_j^n]\mathbf{H}^{-1} = \mathbf{H}[\delta_{ij}\lambda_j^n]\left[\delta_{ij}\frac{\omega_j}{(h_{0j})^2}\right]\mathbf{H}'[\delta_{ij}\sigma_j], \tag{67}$$

where λ_j is given by (66). If we form the (i, k)-entry of $\boldsymbol{\pi}^n$, then we get (65).

In particular, if follows from (65) that

$$p_{00}^{(n)} = \sum_{j\in I} \omega_j\lambda_j^n. \tag{68}$$

We note that

$$\sum_{j\in I} \omega_j = 1 \tag{69}$$

and

$$\omega_0 = \frac{1}{\sigma}. \tag{70}$$

Remark 1. If the polytope $\mathfrak{P}$ has central symmetry, then (22) holds and this implies that

$$\mathbf{A}_{m-\nu} = \mathbf{A}_\nu\mathbf{T} = \mathbf{T}\mathbf{A}_\nu \tag{71}$$

for $\nu \in I$, where

$$\mathbf{T} = [\delta_{i,m-j}]_{i,j\in I}. \tag{72}$$

In this case $\mathbf{H}$ satisfies the matrix equation

$$\mathbf{TH} = \mathbf{HU}, \tag{73}$$

where

$$\mathbf{U} = [\delta_{ij}\varepsilon_j]_{i,j\in I} \tag{74}$$

and

$$\varepsilon_j = \begin{cases} 1 & \text{for } 0 \leqq j \leqq \dfrac{m}{2}, \\ -1 & \text{for } \dfrac{m}{2} < j \leqq m. \end{cases} \tag{75}$$

By (71) and (73) we can prove that

$$\mathbf{\Lambda}_{m-\nu} = \mathbf{\Lambda}_\nu \mathbf{U} \tag{76}$$

for $\nu \in I$.

The matrices $\mathbf{T}$ and $\mathbf{U}$ satisfy the equations

$$\mathbf{T}^2 = \mathbf{U}^2 = \mathbf{I}. \tag{77}$$

Remark 2. In addition to the Euclidean distance between two vertices $\mathbf{x}_r$ and $\mathbf{x}_s$, $\|\mathbf{x}_s - \mathbf{x}_r\|$, we can define another distance, $D(\mathbf{x}_r, \mathbf{x}_s)$, as the minimum number of edges in the paths connecting $\mathbf{x}_r$ and $\mathbf{x}_s$. In a regular polytope two vertices are connected by an edge if and only if their distance is d_1.

For every regular polytope, except the four-dimensional 24-cells, 120-cells and 600-cells, $D(\mathbf{x}_r, \mathbf{x}_s) = j$ if and only if $\|\mathbf{x}_s - \mathbf{x}_r\| = d_j$. Even for the 24-cell and for the 600-cell $D(\mathbf{x}_r, \mathbf{x}_s)$ is uniquely determined by $\|\mathbf{x}_s - \mathbf{x}_r\|$.

Accordingly, if, as an alternative to (5), we assume that the transition probability $\mathbf{P}\{\mathbf{v}_n = \boldsymbol{x}_s | \mathbf{v}_{n-1} = \mathbf{x}_r\}$ depends only on $D(\mathbf{x}_r, \mathbf{x}_s)$, then $p(n)$ can be determined by the same formula as in the case of (5).

If $\mathbf{x}_r \in S_j$, then we write

$$D(\mathbf{x}_r, \mathbf{x}_0) = D_j, \tag{78}$$

whereas $\|\mathbf{x}_r - \mathbf{x}_0\| = d_j$.

In the rest of the paper we shall give the matrices $\mathbf{A}_\nu$ ($\nu = 0, 1, \cdots, m$), $\mathbf{H}$ and $\mathbf{\Lambda}$ for each particular polytope. If we know either $\mathbf{H}$ or $\mathbf{\Lambda}$, the transition probabilities $p_{ik}^{(n)}$ are determined by (65). In each case we shall choose $\lambda_{0\nu} = \sigma_\nu$ for $\nu \in I$; however, this is not essential.

4. Regular polytopes in general. Regular polytopes in two dimensions (regular polygons) and in three dimensions (Platonic solids) have been known from ancient times. Four- and higher-dimensional polytopes were discovered by L. Schläfli [12] before 1853.

Denote by $N_0, N_1, N_2, N_3, \cdots$ the numbers of vertices, edges, faces, cells, $\cdots$ of a polytope. In generalizing Euler's formula for polyhedra, L. Schläfli proved that for an r-dimensional simply connected polytope we have

$$\sum_{i=0}^{r-1} (-1)^i N_i = 1 - (-1)^r. \tag{79}$$

We define a vertex figure of an r-dimensional regular polytope as an $(r-1)$-dimensional regular polytope whose vertices are the midpoints of all the edges of the r-dimensional regular polytope which originate in a given vertex.

We characterize regular polytopes by the Schläfli symbols. The Schläfli symbol of a regular p-gon is $\{p\}$. A regular polyhedron which has p-gonal faces, q at each vertex, is characterized by the Schläfli symbol $\{p, q\}$. A four-dimensional regular polytope which has Schläfli symbol $\{p, q, r\}$ has cells of type $\{p, q\}$ and vertex figures $\{q, r\}$. Similarly, a general regular polytope $\{p, q, \cdots, v, w\}$ has cells $\{p, q, \cdots, v\}$ and vertex figures $\{q, \cdots, v, w\}$.

Tables 1, 2 and 3 show all the regular polytopes in three, four, and higher dimensions. For the theory of regular polytopes we refer to H. S. M. Coxeter [1], L. Schläfli [12], P. H. Schoute [11] and D. M. Y. Sommerville [13].

TABLE 1
Regular polyhedra in three dimensions

Polyhedron	Schläfli symbol	N_0	N_1	N_2
Tetrahedron	{3, 3}	4	6	4
Octahedron	{3, 4}	6	12	8
Cube	{4, 3}	8	12	6
Icosahedron	{3, 5}	12	30	20
Dodecahedron	{5, 3}	20	30	12

TABLE 2
Regular polytopes in four dimensions

Polytope	Schläfli symbol	N_0	N_1	N_2	N_3
5-cell	{3, 3, 3}	5	10	10	5
16-cell	{3, 3, 4}	8	24	32	16
8-cell	{4, 3, 3}	16	32	24	8
24-cell	{3, 4, 3}	24	96	96	24
600-cell	{3, 3, 5}	120	720	1200	600
120-cell	{5, 3, 3}	600	1200	720	120

TABLE 3
Regular polytopes in r dimensions ($r \geqq 5$)

Polytope	Schläfli symbol	N_0	$\cdots$	N_j	$\cdots$	N_{r-1}
Regular simplex	$\{3, \cdots, 3\}$	$r+1$	$\cdots$	$\binom{r+1}{j+1}$	$\cdots$	$r+1$
Cross polytope	$\{3, \cdots, 3, 4\}$	$2r$	$\cdots$	$\binom{r}{j+1}2^{j+1}$	$\cdots$	2^r
Measure polytope	$\{4, 3, \cdots, 3\}$	2^r	$\cdots$	$\binom{r}{j}2^{r-j}$	$\cdots$	$2r$

5. Regular polygons. Let $\mathfrak{P}$ be a regular t-gon with circumradius $\rho = 1$. We can choose

$$\mathbf{x}_r = \left(\cos\left(\frac{2\pi r}{t}\right), \sin\left(\frac{2\pi r}{t}\right)\right) \tag{80}$$

$(r = 0, 1, \cdots, t-1)$ as the vertices of $\mathfrak{P}$. Then $\sigma = t$, $m = [t/2]$, $S_0 = \{\mathbf{x}_0\}$ and $S_j = \{\mathbf{x}_j, \mathbf{x}_{t-j}\}$ for $j = 1, 2, \cdots, m$. Now we have $\sigma_0 = 1$, $\sigma_j = 2$ if $1 \leqq j < t/2$ and $\sigma_j = 1$ if $j = t/2$ and t is even. The probabilities $p_0, p_1, \cdots, p_m$ satisfy

$$p_0 + 2p_1 + \cdots + 2p_{m-1} + p_m = 1 \tag{81}$$

if t is even, and

$$p_0 + 2p_1 + \cdots + 2p_{m-1} + 2p_m = 1 \tag{82}$$

if t is odd.

For any event A, let $\delta(A) = 1$ if A occurs and $\delta(A) = 0$ if A does not occur. Then we can write that

$$a_{ij\nu} = \delta(j = |i - \nu|) + \delta\left(j = \frac{t}{2} - \left|\frac{t}{2} - i - \nu\right|\right) \tag{83}$$

for $1 \leqq \nu \leqq m-1$, $a_{ij0} = \delta_{ij}$ and $a_{ijm} = \delta_{i,m-j}$.

The eigenvalues of $\mathbf{A}_\nu$ are

$$\lambda_{j\nu} = \sigma_\nu \cos\left(\frac{2\pi j\nu}{t}\right) \tag{84}$$

for $j \in I$ and $\nu \in I$. Now we can choose $h_{ij} = \lambda_{ij}$ and we have $\omega_j = \sigma_j/t$ for $j \in I$.

The transition probabilities $p_{ik}^{(n)}$ are given by (65), and in particular we have

$$tp_{00}^{(n)} = \sum_{j=0}^{m} \sigma_j \lambda_j^n, \tag{85}$$

where λ_j is defined by (45).

Now the Markov chain $\{\mathbf{v}_n; n = 0, 1, 2, \cdots\}$ has state space $(0, 1, \cdots, t-1)$ and transition probability matrix $[p_{i+j}]$, where $p_{t+\nu} = p_\nu$ $(\nu = 0, 1, 2, \cdots)$ and $p_\nu = p_{t-\nu}$ $(t/2 < \nu \leqq t)$. The transition probability matrix is cyclic and has the following Jordan decomposition:

$$[p_{i+j}]_{i,j} = \frac{1}{t}[\varepsilon_j^i]_{i,j}[\delta_{ij}\lambda_j][\varepsilon_j^{-k}]_{j,k}, \tag{86}$$

where

$$\varepsilon_j = \exp\left(\frac{2\pi\sqrt{-1}\,j}{t}\right) \tag{87}$$

and

$$\lambda_j = \sum_{\nu=0}^{t-1} p_\nu \varepsilon_j^\nu. \tag{88}$$

6. Regular simplexes. In two dimensions a regular simplex is an equilateral triangle, in three dimensions a tetrahedron, and in four dimensions a 5-cell. In r dimensions $(r \geqq 2)$ we can choose $\mathbf{x}_0 = (1, 1, \cdots, 1)$ and $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_r$ as the r cyclic permutations of $(1, -1, \cdots, -1)$ for the vertices of a regular simplex. Then $\sigma_0 = r+1$, $m = 1$, $S_0 = \{\mathbf{x}_0\}$ and $S_1 = \{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$. Now we have $\sigma_0 = 1$, $\sigma_1 = r$ and

$$\mathbf{A}_1 = \begin{bmatrix} 0 & r \\ 1 & r-1 \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} 1 & r \\ 1 & -1 \end{bmatrix}, \quad \mathbf{\Lambda} = \begin{bmatrix} 1 & r \\ 1 & -1 \end{bmatrix}. \tag{89}$$

Since $\omega_0 = 1/(r+1)$ and $\omega_1 = r/(r+1)$, therefore

$$p_{00}^{(n)} = \frac{1}{r+1} + \frac{r}{r+1}(p_0 - p_1)^n \tag{90}$$

for $n \geqq 0$, where $p_0 + rp_1 = 1$.

7. Regular cross polytopes. In two dimensions a regular cross polytope is a square, in three dimensions an octahedron, and in four dimensions a 16-cell. In r dimensions ($r \geqq 2$) we can choose $\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_{2r-1}$ as the $2r$ permutations of $(\pm 1, 0, \cdots, 0)$ for the vertices of a cross polytope. Then $\sigma = 2r$, $m = 2$, $\sigma_0 = 1$, $\sigma_1 = 2(r-1)$, $\sigma_2 = 1$ and

$$\mathbf{A}_1 = \begin{bmatrix} 0 & 2(r-1) & 0 \\ 1 & 2(r-2) & 1 \\ 0 & 2(r-1) & 0 \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} 1 & (r-1) & 1 \\ 1 & -1 & 0 \\ 1 & (r-1) & -1 \end{bmatrix}, \quad \mathbf{\Lambda} = \begin{bmatrix} 1 & 2(r-1) & 1 \\ 1 & -2 & 1 \\ 1 & 0 & -1 \end{bmatrix}. \tag{91}$$

Since $\omega_0 = 1/(2r)$, $\omega_1 = (r-1)/(2r)$ and $\omega_2 = \frac{1}{2}$, therefore

$$p_{00}^{(n)} = \frac{1}{2r} + \frac{(r-1)}{2r}(p_0 - 2p_1 + p_2)^n + \tfrac{1}{2}(p_0 - p_2)^n \tag{92}$$

for $n \geqq 0$, where $p_0 + 2(r-1)p_1 + p_2 = 1$.

8. Measure polytopes. In two dimensions a measure polytope is a square, in three dimensions a cube, and in four dimensions an 8-cell. In r dimensions ($r \geqq 2$) a measure polytope has 2^r vertices, say,

$$\{(\alpha_1, \alpha_2, \cdots, \alpha_r): \alpha_i = 1 \text{ or } -1 \text{ for all } i = 1, 2, \cdots, r\}. \tag{93}$$

Then $\sigma = 2^r$, $m = r$ and $\sigma_j = \binom{m}{j}$ for $j = 0, 1, \cdots, m$. Now

$$\lambda_{ij} = \sum_{\nu=0}^{j} (-1)^\nu \binom{i}{\nu}\binom{m-i}{j-\nu} \tag{94}$$

for $0 \leqq i \leqq m$ and $0 \leqq j \leqq m$. We can write down also that

$$\sum_{j=0}^{m} \lambda_{ij} z^j = (1-z)^i (1+z)^{m-i} \tag{95}$$

for $0 \leqq i \leqq m$. Now we can choose $h_{ij} = \lambda_{ij}$, and then

$$\omega_j = \binom{m}{j}\frac{1}{2^m}. \tag{96}$$

In this case

$$p_{ik}^{(n)} = \frac{1}{2^m} \sum_{j=0}^{m} h_{ij} h_{jk} \lambda_j^n, \tag{97}$$

where

$$\lambda_j = \sum_{\nu=0}^{m} h_{j\nu} p_\nu \tag{98}$$

for $j = 0, 1, \cdots, m$ and $p_0, p_1, \cdots, p_m$ satisfy the requirement

$$\sum_{\nu=0}^{m} \binom{m}{\nu} p_\nu = 1. \tag{99}$$

In the particular case where $p_1 = 1/m$ and $p_\nu = 0$ for $\nu \neq 1$, the above results can be deduced from some results of M. Kac [3], [4] for the Ehrenfest urn problem. See also E. Schrödinger [10], F. G. Hess [2], M. J. Klein [5], L. Takács [15] and G. Letac and L. Takács [8].

Now we have

$$a_{i,j,i+j-2\nu} = \binom{i}{\nu}\binom{m-i}{j-\nu}. \tag{100}$$

9. The icosahedron. The icosahedron has $\sigma = 12$ vertices which may be chosen as the cyclic permutations of $(1, 0, \tau)$ with all changes of sign. Here and in the rest of the paper

$$\tau = \frac{1+\sqrt{5}}{2}. \tag{101}$$

In this case $m = 3$, $\mathbf{A}_0 = \mathbf{I}$, $\mathbf{A}_1$ is given below, $\mathbf{A}_2 = \mathbf{A}_1\mathbf{T}$ and $\mathbf{A}_3 = \mathbf{T}$, where $\mathbf{T}$ is defined by (72). We have

$$\mathbf{A}_1 = \begin{bmatrix} \cdot & 5 & \cdot & \cdot \\ 1 & 2 & 2 & \cdot \\ \cdot & 2 & 2 & 1 \\ \cdot & \cdot & 5 & 0 \end{bmatrix}, \tag{102}$$

where the dot means 0. Tables 4, 5 and 6 furnish all the data needed to find $p_{ik}^{(n)}$.

TABLE 4
The icosahedron

j	0	1	2	3
σ_j	1	5	5	1
$12\omega_j$	1	5	3	3

TABLE 5
h_{ij}

i \ j	0	1	2	3
0	1	5	$\sqrt{5}$	$\sqrt{5}$
1	1	−1	1	−1
2	1	−1	−1	1
3	1	5	$-\sqrt{5}$	$-\sqrt{5}$

TABLE 6
$\lambda_{j\nu}$

j \ ν	0	1	2	3
0	1	5	5	1
1	1	−1	−1	1
2	1	$\sqrt{5}$	$-\sqrt{5}$	−1
3	1	$-\sqrt{5}$	$\sqrt{5}$	−1

The n-step transition probabilities are given by (65). In particular, we have

$$12p_{00}^{(n)} = \lambda_0^n + 5\lambda_1^n + 3\lambda_2^n + 3\lambda_3^n \tag{103}$$

for $n \geqq 0$.

10. The dodecahedron. The dodecahedron has $\sigma = 20$ vertices which may be chosen as $(\pm 1, \pm 1, \pm 1)$ and the cyclic permutations of $(0, \tau^{-1}, \tau)$ with all changes of sign. Here τ is defined by (101). In this case $m = 5$, $\mathbf{A}_0 = \mathbf{I}$,

$$(104)\quad \mathbf{A}_1 = \begin{bmatrix} \cdot & 3 & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & 2 & \cdot & \cdot & \cdot \\ \cdot & 1 & 1 & 1 & \cdot & \cdot \\ \cdot & \cdot & 1 & 1 & 1 & \cdot \\ \cdot & \cdot & \cdot & 2 & \cdot & 1 \\ \cdot & \cdot & \cdot & \cdot & 3 & \cdot \end{bmatrix}, \qquad \mathbf{A}_2 = \begin{bmatrix} \cdot & \cdot & 6 & \cdot & \cdot & \cdot \\ \cdot & 2 & 2 & 2 & \cdot & \cdot \\ 1 & 1 & 1 & 2 & 1 & \cdot \\ \cdot & 1 & 2 & 1 & 1 & 1 \\ \cdot & \cdot & 2 & 2 & 2 & \cdot \\ \cdot & \cdot & \cdot & 6 & \cdot & \cdot \end{bmatrix}.$$

Now $p_{ik}^{(n)}$ is given by (65), and in particular we have

$$(105)\qquad 20p_{00}^{(n)} = \lambda_0^n + 4\lambda_1^n + 5\lambda_2^n + 4\lambda_3^n + 3\lambda_4^n + 3\lambda_5^n$$

for $n \geqq 0$. See G. Letac and L. Takács [7], where (105) is determined by another method.

11. The 24-cell. The 24-cell has $\sigma = 24$ vertices and $m = 4$. See Table 7 for the sections of a 24-cell with circumradius 2.

TABLE 7
The 24-cell

j	S_j	σ_j	$24\omega_j$	d_j	D_j
0	(0, 0, 0, 2)	1	1	0	0
1	(1, 1, 1, 1)	8	2	2	1
2	(2, 0, 0, 0)	6	9	$2\sqrt{2}$	2
3	(1, 1, 1, −1)	8	8	$2\sqrt{3}$	2
4	(0, 0, 0, −2)	1	4	4	3

In each section only one representative vertex is displayed. To obtain all the vertices in S_j we should equip the first three coordinates of the displayed vertex by the signs $\pm$ and form all permutations of the first three coordinates.

Now $\mathbf{A}_0 = \mathbf{I}$,

$$(106)\qquad \mathbf{A}_1 = \begin{bmatrix} \cdot & 8 & \cdot & \cdot & \cdot \\ 1 & 3 & 3 & 1 & \cdot \\ \cdot & 4 & \cdot & 4 & \cdot \\ \cdot & 1 & 3 & 3 & 1 \\ \cdot & \cdot & \cdot & 8 & \cdot \end{bmatrix}, \qquad \mathbf{A}_2 = \begin{bmatrix} \cdot & \cdot & 6 & \cdot & \cdot \\ \cdot & 3 & \cdot & 3 & \cdot \\ 1 & \cdot & 4 & \cdot & 1 \\ \cdot & 3 & \cdot & 3 & \cdot \\ \cdot & \cdot & 6 & \cdot & \cdot \end{bmatrix},$$

$\mathbf{A}_3 = \mathbf{A}_1\mathbf{T}$ and $\mathbf{A}_4 = \mathbf{T}$, where $\mathbf{T}$ is defined by (72).

The elements of the matrices $\mathbf{H}$ and Λ are given by Tables 8 and 9.

TABLE 8
h_{ij}

i \ j	0	1	2	3	4
0	1	2	3	4	2
1	1	−1	0	−1	1
2	1	2	−1	0	0
3	1	−1	0	1	−1
4	1	2	3	−4	−2

TABLE 9
$\lambda_{j\nu}$

j \ ν	0	1	2	3	4
0	1	8	6	8	1
1	1	−4	6	−4	1
2	1	0	−2	0	1
3	1	−2	0	2	−1
4	1	4	0	−4	−1

Now $p_{ik}^{(n)}$ is given by (65), and in particular we have

$$24p_{00}^{(n)} = \lambda_0^n + 2\lambda_1^n + 9\lambda_2^n + 8\lambda_3^n + 4\lambda_4^n \tag{107}$$

for $n \geqq 0$.

12. The 600-cell. A 600-cell has $\sigma = 120$ vertices. A 600-cell with circumradius $\rho = 2$ and center $(0, 0, 0, 0)$ contains the following vertices: the 8 permutations of $(\pm 2, 0, 0, 0)$, the 16 permutations of $(\pm 1, \pm 1, \pm 1, \pm 1)$ and the 96 even permutations of $(\pm\tau, \pm 1, \pm\tau^{-1}, 0)$, where τ is defined by (101). Now $m = 8$. See Table 10 for the sections of this polytope. In each section only some representative vertices are displayed. To obtain all the vertices in S_j we should equip the first three coordinates of the displayed vertices by the signs $\pm$ and form all permutations of the first three coordinates.

TABLE 10
The 600-cell

j	S_j	σ_j	$120\omega_j$	d_j^2	D_j
0	$(0, 0, 0, 2)$	1	1	0	0
1	$(1, 0, \tau^{-1}, \tau)$	12	16	$6-2\sqrt{5}$	1
2	$(1, 1, 1, 1)$ $(\tau, \tau^{-1}, 0, 1)$	20	9	4	2
3	$(\tau, 0, 1, \tau^{-1})$	12	9	$10-2\sqrt{5}$	2
4	$(2, 0, 0, 0)$ $(\tau, 1, \tau^{-1}, 0)$	30	25	8	3
5	$(\tau, 0, 1, \tau^{-1})$	12	16	$6+2\sqrt{5}$	3
6	$(1, 1, 1, 1)$ $(\tau, \tau^{-1}, 0, -1)$	20	36	12	4
7	$(1, 0, \tau^{-1}, -\tau)$	12	4	$10+2\sqrt{5}$	4
8	$(0, 0, 0, -1)$	1	4	16	5

Now $\mathbf{A}_0 = \mathbf{I}$, $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, $\mathbf{A}_4$, are given in Tables 11, 12, 13 and 14, and $\mathbf{A}_5 = \mathbf{A}_3\mathbf{T}$, $\mathbf{A}_6 = \mathbf{A}_2\mathbf{T}$, $\mathbf{A}_7 = \mathbf{A}_1\mathbf{T}$ and $\mathbf{A}_8 = \mathbf{T}$.

TABLE 11
a_{ij1}

i \ j	0	1	2	3	4	5	6	7	8
0		12							
1	1	5	5	1					
2		3	3	3	3				
3		1	5		5	1			
4			2	2	4	2	2		
5				1	5		5	1	
6					3	3	3	3	
7						1	5	5	1
8								12	

TABLE 12
a_{ij2}

i \ j	0	1	2	3	4	5	6	7	8
0			20						
1		5	5	5	5				
2	1	3	6		6	3	1		
3		5		5	5		5		
4		2	4	2	4	2	4	2	
5			5		5	5		5	
6			1	3	6		6	3	1
7					5	5	5	5	
8							20		

TABLE 13
a_{ij3}

i \ j	0	1	2	3	4	5	6	7	8
0				12					
1		1	5		5	1			
2		3		3	3		3		
3	1		5			5		1	
4		2	2		4		2	2	
5		1		5			5		1
6			3		3	3		3	
7				1	5		5	1	
8						12			

TABLE 14
a_{ij4}

i \ j	0	1	2	3	4	5	6	7	8
0					30				
1			5	5	10	5	5		
2		3	6	3	6	3	6	3	
3		5	5		10		5	5	
4	1	4	4	4	4	4	4	4	1
5		5	5		10		5	5	
6		3	6	3	6	3	6	3	
7			5	5	10	5	5		
8					30				

The elements of the matrices $\mathbf{H}$ and Λ are given in Tables 15 and 16.

TABLE 15
h_{ij}

i \ j	0	1	2	3	4	5	6	7	8
0	1	4	3	3	5	4	6	2	2
1	1	-1	τ	$-\tau^{-1}$	0	1	-1	τ	$-\tau^{-1}$
2	1	1	0	0	-1	-1	0	1	1
3	1	-1	$-\tau^{-1}$	τ	0	-1	1	τ^{-1}	$-\tau$
4	1	0	-1	-1	1	0	0	0	0
5	1	-1	$-\tau^{-1}$	τ	0	1	-1	$-\tau^{-1}$	τ
6	1	1	0	0	-1	1	0	-1	-1
7	1	-1	τ	$-\tau^{-1}$	0	-1	1	$-\tau$	τ^{-1}
8	1	4	3	3	5	-4	-6	-2	-2

TABLE 16
$\lambda_{j\nu}$

j \ ν	0	1	2	3	4	5	6	7	8
0	1	12	20	12	30	12	20	12	1
1	1	-3	5	-3	0	-3	5	-3	1
2	1	4τ	0	$-4\tau^{-1}$	-10	$-4\tau^{-1}$	0	4τ	1
3	1	$-4\tau^{-1}$	0	4τ	-10	4τ	0	$-4\tau^{-1}$	1
4	1	0	-4	0	6	0	-4	0	1
5	1	3	-5	-3	0	3	5	-3	-1
6	1	-2	0	2	0	-2	0	2	-1
7	1	6τ	10	$6\tau^{-1}$	0	$-6\tau^{-1}$	-10	-6τ	-1
8	1	$-6\tau^{-1}$	10	-6τ	0	6τ	-10	$6\tau^{-1}$	-1

The n-step transition probabilities are given by (65). In particular, we have

$$120p_{00}^{(n)} = \lambda_0^n + 16\lambda_1^n + 9\lambda_2^n + 9\lambda_3^n + 25\lambda_4^n + 16\lambda_5^n + 36\lambda_6^n + 4\lambda_7^n + 4\lambda_8^n \tag{108}$$

for $n \geqq 0$. See also G. Letac and L. Takács [9].

13. The 120-cell. A 120-cell has 600 vertices. Now the state space of the Markov chain $\{\mathbf{v}_n;\ n = 0, 1, 2, \cdots\}$ consists of 600 states. If we define the sequence of random variables $\{\xi_n;\ n = 0, 1, 2, \cdots\}$ such that $\xi_n = j$ whenever $\mathbf{v}_n \in S_j$, where S_j is defined by (2), then the state space of $\{\xi_n;\ n = 0, 1, 2, \cdots\}$ consists of 31 states, but $\{\xi_n;\ n = 0, 1, 2, \cdots\}$ is not a Markov chain. If we want to restore the Markov property of $\{\xi_n;\ n = 0, 1, 2, \cdots\}$, then some of the sections S_j $(j = 0, 1, \cdots, 30)$ should be subdivided into two or three disjoint subsets. Proceeding in this way we arrive at a homogeneous Markov chain having 45 states. We shall study this Markov chain in a subsequent paper. Here we would like to mention only that if $p_1 = \frac{1}{4}$ and $p_\nu = 0$ otherwise, that is, if in each flight the traveler visits one of the four neighboring vertices of the starting vertex with probability $p_1 = \frac{1}{4}$, then the probability $p(n)$ that the traveler returns to the initial

position in n steps is given by

$$
\begin{aligned}
600p(n)4^n = {} & 4^n + 8(-2)^n + 8(-1)^n + 18.0^n + 40 + 9\left\{\left(\frac{5+\sqrt{5}}{2}\right)^n + \left(\frac{5-\sqrt{5}}{2}\right)^n\right\} \\
& + 24\{(\sqrt{5})^n + (-\sqrt{5})^n\} + 30(-1)^n\left\{\left(\frac{3+\sqrt{5}}{2}\right)^n + \left(\frac{3-\sqrt{5}}{2}\right)^n\right\} \\
& + 16\left\{\left(\frac{\sqrt{21}-1}{2}\right)^n + (-1)^n\left(\frac{\sqrt{21}+1}{2}\right)^n\right\} + 4\left\{\left(\frac{3\sqrt{5}+1}{2}\right)^n + (-1)^n\left(\frac{3\sqrt{5}-1}{2}\right)^n\right\} \\
& + 16\left\{\left(\frac{\sqrt{13}+3}{2}\right)^n + (-1)^n\left(\frac{\sqrt{13}-3}{2}\right)^n\right\} + 24\left\{\left(\frac{\sqrt{5}+1}{2}\right)^n + (-1)^n\left(\frac{\sqrt{5}-1}{2}\right)^n\right\} \\
& + 48\{(\sqrt{2}-1)^n + (-1)^n(\sqrt{2}+1)^n\} \\
& + 25\left\{\left(\frac{1}{3} + \frac{2\sqrt{22}}{3}\cos\frac{\varphi}{3}\right)^n\right. \\
& \qquad \left. + \left(\frac{1}{3} + \frac{2\sqrt{22}}{3}\cos\frac{(\varphi+2\pi)}{3}\right)^n + \left(\frac{1}{3} + \frac{2\sqrt{22}}{3}\cos\frac{(\varphi+4\pi)}{3}\right)^n\right\} \\
& + 36\left\{\left(\frac{1}{3} + \frac{2\sqrt{22}}{3}\cos\frac{\psi}{3}\right)^n + \left(\frac{1}{3} + \frac{2\sqrt{22}}{3}\cos\frac{(\psi+2\pi)}{3}\right)^n\right. \\
& \qquad \left. + \left(\frac{1}{3} + \frac{2\sqrt{22}}{3}\cos\frac{(\psi+4\pi)}{3}\right)^n\right\},
\end{aligned}
\tag{109}
$$

where

$$
\cos\varphi = -\frac{43}{44\sqrt{22}} \quad \text{and} \quad \cos\Psi = -\frac{151}{44\sqrt{22}}. \tag{110}
$$

Note added in proof. After formula (104) the following should be added: $\mathbf{A}_3 = \mathbf{A}_2\mathbf{T}$, $\mathbf{A}_4 = \mathbf{A}_1\mathbf{T}$, and $\mathbf{A}_5 = \mathbf{T}$, where $\mathbf{T}$ is defined by (72). The following tables furnish all the data needed to find $p_{ik}^{(n)}$.

TABLE 17
The dodecahedron

j	0	1	2	3	4	5
σ_j	1	3	6	6	3	1
$20\omega_j$	1	4	5	4	3	3

TABLE 18
h_{ij}

i \ j	0	1	2	3	4	5
0	1	6	3	2	3	3
1	1	−4	1	0	$-\sqrt{5}$	$\sqrt{5}$
2	1	1	−1	−1	1	1
3	1	1	−1	1	−1	−1
4	1	−4	1	0	$\sqrt{5}$	$-\sqrt{5}$
5	1	6	3	−2	−3	−3

TABLE 19
$\lambda_{j\nu}$

j \ ν	0	1	2	3	4	5
0	1	3	6	6	3	1
1	1	−2	1	1	−2	1
2	1	1	−2	−2	1	1
3	1	0	−3	3	0	−1
4	1	$-\sqrt{5}$	2	−2	$\sqrt{5}$	−1
5	1	$\sqrt{5}$	2	−2	$-\sqrt{5}$	−1

REFERENCES

[1] H. S. M. COXETER, *Regular Polytopes*, 2nd edition, Macmillan, New York, 1963.

[2] F. G. HESS, *Alternative solution to the Ehrenfest problem*, Amer. Math. Monthly, 61 (1954), pp. 323–327.

[3] M. KAC, *Random walk and the theory of Brownian motion*, Amer. Math. Monthly, 54 (1947), pp. 369–391; reprinted in The Chauvenet Papers, vol. I, J. C. Abbott, ed., Mathematical Association of America, Washington, DC, 1978, pp. 253–275.

[4] ———, *Appendix to Random walk and the theory of Brownian motion*, in The Chauvenet Papers, vol. I, J. C. Abbott, ed., Mathematical Association of America, Washington, DC, 1978, pp. 276–277.

[5] M. J. KLEIN, *Generalization of the Ehrenfest urn model*, Phys. Rev., 103 (1956), pp. 17–20.

[6] G. LETAC, *Problem* 6149, Amer. Math. Monthly, 84 (1977), p. 301.

[7] G. LETAC AND L. TAKÁCS, *Random walks on a dodecahedron*, J. Appl. Prob., 17 (1980), pp. 373–384.

[8] ———, *Random walks on an m-dimensional cube*, J. Reine Angew Math., 310 (1979), pp. 187–195.

[9] ———, *Random walks on a* 600*-cell*, this Journal, 1 (1980), pp. 114–120.

[10] E. SCHRÖDINGER, *Quantisierung als Eigenwertproblem III*, Annal. Physik, 80 (1926), pp. 437–490; English translation in E. Schrödinger, Collected Papers on Wave Mechanics, Chelsea, New York, 1978, pp. 62–101.

[11] P. H. SCHOUTE, *Mehrdimensionale Geometrie. II. Teil. Die Polytope*, Sammlung Schubert XXXVI, G. J. Göschen'sche Verlagshandlung, Leipzig, 1905.

[12] L. SCHLÄFLI, *Theorie der vielfachen Kontinuität.* Denkschriften der Schweizerischen naturforschenden Gesellschaft, 38 (1901), pp. 1–237; also in Ludwig Schläfli (1814–1895) Gesammelte Mathematische Abhandlungen, Band I, Verlag-Birkhäuser, Basel, 1950, pp. 167–387.

[13] D. M. Y. SOMMERVILLE, *An Introduction to the Geometry of N Dimensions*, Methuen, London, 1929; reprinted by Dover, New York, 1958.

[14] L. TAKÁCS, *Walk on the edges of a dodecahedron* (solution of Problem 6149), Amer. Math. Monthly, 86 (1979), pp. 61–63.

[15] ———, *On an urn problem of Paul and Tatiana Ehrenfest*, Math. Proc. Cambridge Phil. Soc., 86 (1979), pp. 127–130.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 2, June 1981

0196-5212/81/0202-0011 $01.00/0

ADJACENCY ON THE POSTMAN POLYHEDRON*

RICK GILES†

Abstract. Let $G = (V, E)$ be a loopless, undirected graph and $C \subseteq V$ have even cardinality. A *postman set* is a subset $J \subseteq E$ such that for every node $v \in V$, the number of edges of J incident to v is odd if and only if $v \in C$. The *postman polyhedron* $P(G)$ is the sum of the convex hull of all incidence vectors of postman sets and the nonnegative orthant $\mathbb{R}_+^E$. We give a simple characterization of adjacency for vertices of $P(G)$. An upper bound on the distance between two vertices, and hence the diameter of $P(G)$, is given.

Let $G = (V, E)$ be a loopless undirected graph with *node* set V and *edge* set E, and let $C \subseteq V$ be given, with $|C|$ even. Subset $J \subseteq E$ is called a *postman set* if for all $v \in V$, $|J \cap \delta(v)| \equiv 1 \pmod 2$ if and only if $v \in C$, where $\delta(v)$ denotes the set of edges incident to v. It may be the case that G has no postman set. However, it is not difficult to see that G has a postman set if and only if each component of G contains an even number of nodes of C, and henceforth we assume that G has a postman set.

The following two important examples of postman sets are obtained by making appropriate choices of C. First, if $|C| = 2$, say $C = \{s, t\}$, then J is a minimal postman set if and only if J is the edge set of an $s - t$ path. Second, if $C = \{v \in V: v \text{ has odd degree}\}$, then a postman set corresponds to a set of edges such that if an additional copy of each edge in the set is made, then the resulting graph is Eulerian. We refer the reader to Edmonds and Johnson [4] for the details of this correspondence.

Edmonds and Johnson [4] describe an efficient algorithm for the problem of finding a minimum weight-sum postman set. Their work includes the description, as the solution set of a linear system, of the following unbounded polyhedron associated with postman sets. Let $\mathbb{R}^E$ denote the set of real-valued vectors $(x_j: j \in E)$. For $J \subseteq E$, x^J denotes the incidence vector of J. We define the *postman polyhedron*, $P(G) \equiv \text{conv}\{x^J: J \text{ is a postman set}\} + \mathbb{R}_+^E$. Thus the vertices of $P(G)$ are the incidence vectors of minimal postman sets and, given a vector $c \in \mathbb{R}_+^E$, the Edmonds–Johnson algorithm is a good algorithm for finding a vertex of $P(G)$ which minimizes cx over $x \in P(G)$.

Here we study the adjacency relation between vertices of $P(G)$. Two vertices w and x of a polyhedron P are said to be *adjacent* if the line segment $[w, x]$ is an edge of P, i.e., no point of $[w, x]$ is a convex combination of points of $P - [w, x]$. Two postman sets are *adjacent* if their respective incidence vectors are adjacent on $P(G)$. Before giving a characterization of adjacency for postman sets, we establish three properties of postman sets.

PROPERTY 1. *Let $J \subseteq E$ and let P be the edge set of a polygon. Then J is a postman set if and only if $J \triangle P$ is a postman set, where $\triangle$ denotes symmetric difference.*

Proof. If $J \subseteq E$, P is the edge set of a polygon and $v \in V$, then

$$|J \cap \delta(v)| \equiv |(J \triangle P) \cap \delta(v)| \pmod 2. \qquad \square$$

PROPERTY 2. *If J and K are two different postman sets, then $J \triangle K$ contains the edge set of a polygon.*

Property 2 follows from the observation that if J and K are two different postman sets, then for every node v, $|(J \triangle K) \cap \delta(v)|$ is even, and so $J \triangle K$ contains the edge set

* Received by the editors October 15, 1979, and in final form November 20, 1980. This research was supported in part by the National Science Foundation under grant MCS 78-01982, and by Sonderforschungsbereich 21(DFG), Institut für Ökonometrie und Operations Research, Universität Bonn, Bonn, West Germany.

† Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506.

of a polygon. Notice that Properties 1 and 2 imply that a postman set is minimal if and only if it does not contain the edge set of a polygon.

PROPERTY 3. *Let J and K be two minimal postman sets and let $E = J \cup K$. Then $J \cup K$ contains the edge set of only one polygon if and only if J and K are the only postman sets.*

Proof. Suppose $J \cup K$ contains the edge sets, say P and Q, of two different polygons. Then, by Property 1, $J \triangle P$ and $J \triangle Q$ are two postman sets, each different from J. If $J \triangle P = K = J \triangle Q$, then $P = J \triangle K = Q$; a contradiction. Hence $J \triangle P$ or $J \triangle Q$ is a third postman set.

Conversely, suppose $J \cup K$ contains the edge set, say P, of only one polygon. If $J \triangle P \neq K$, then $(J \triangle P) \triangle K = (J \triangle K) - P$ contains the edge set of another polygon, a contradiction. Hence $J \triangle P = K$ and $J \triangle K = P$. If L is a third postman set, then, because $L \subseteq J \cup K$ and $J \triangle L$, $K \triangle L$ each contain the edge set of a polygon, $J \triangle L \supseteq P \subseteq K \triangle L$. But then $P - J \subseteq L$ and $P - K = P \cap J \subseteq L$; whence $P \subseteq L$. $L \triangle P \subseteq J \cap K$ is also a postman set; contradicting the minimality of J and of K. Hence there is no third postman set. □

THEOREM 1. *Let J and K be two minimal postman sets. The following are equivalent.*

(i) *J, K are adjacent.*

(ii) *There do not exist two postman sets L, M, different from J, K, such that $L \cap M \subseteq J \cap K$ and $L \cup M \subseteq J \cup K$.*

(iii) *$J \cup K$ contains the edge set of only one polygon.*

Proof. (i) $\Rightarrow$ (ii). Suppose L, M are two postman sets, different from J, K, such that $L \cap M \subseteq J \cap K$ and $L \cup M \subseteq J \cup K$. Where $x = \frac{1}{2}(x^J + x^K)$ and $y = \frac{1}{2}(x^L + x^M)$, we have $x \geqq y$. If $x = y$, then J, K are nonadjacent because x^L and x^M are not on the line segment $[x^J, x^K]$. If $x \neq y$, then $y \notin [x^J, x^K]$ and, because $P(G) = P(G) + \mathbb{R}^E_+$, there is a vector $z \in P(G) - [x^J, x^K]$ such that $\frac{1}{2}(y + z) = x$.

(ii) $\Rightarrow$ (iii). If $J \cup K$ contains the edge sets of two different polygons, then, as was argued in the proof of Property 3, there is the edge set, say P, of a polygon such that $P \subseteq J \cup K$ and $L \equiv J \triangle P$, $M \equiv K \triangle P$ are two postman sets different from J, K. By definition of L and M, $L \cap M \subseteq J \cap K$ and $L \cup M \subseteq J \cup K$.

(iii) $\Rightarrow$ (i). Suppose $J \cup K$ contains the edge set of only one polygon. Let $x \in [x^J, x^K]$. Then $x_j = 0$ for all $j \in E - (J \cup K)$. If x is a convex combination of points of $P(G) - [x^J, x^K]$, then we may assume that one of these points is the incidence vector of a third postman set, which must be contained in $J \cup K$. This contradicts Property 3, so J, K are adjacent. □

Given two vertices w and x of a polyhedron P, there is always a sequence $w = x_0, x_1, \cdots, x_n = x$ such that x_i and x_{i+1} are adjacent on P for $0 \leqq i \leqq n-1$. The *distance* on P from w to x, $d_P(w, x)$, is defined to be the minimum integer n for which a sequence of this type exists. The *diameter* of P is $\max \{d_P(w, x)\colon w, x$ are two vertices of $P\}$.

THEOREM 2. *If J, K are two minimal postman sets, then $d_{P(G)}(x^J, x^K) \leqq \min \{|J - K|, |K - J|\}$.*

Proof. Suppose $|J - K| \leqq |K - J|$. Let $P \subseteq J \cup K$ be the edge set of a polygon such that $P - J$ is minimal. We claim that $L \equiv J \triangle P$ is a minimal postman set. Suppose $Q \subseteq L$ is the edge set of a polygon. Then $Q \subseteq J \cup K$, $Q - J \subseteq P - J$ and, by the minimality of $P - J$, $Q - J = P - J$. It is impossible that $Q = P$ because $P \cap J \neq \varnothing$ and $(P \cap J) \cap L = \varnothing$. But then it is a well-known graph property that for any $j \in (Q - J) \cap (P - J)$, $(P \cup Q) - j$ contains the edge set of a polygon, say R, and $R - J \subsetneq P - J$; contradicting the minimality of $P - J$. Hence L is a minimal postman set. Clearly $J \cup L = J \cup P$. If $J \cup L$ contained the edge set of another polygon, then, as argued above, $P - J$ would not

be minimal. Hence, by Theorem 1, J and L are adjacent. Furthermore, $|L-K|<|J-K|$ and the theorem follows by induction on $|J-K|$. □

The bound in Theorem 2 may be realized. For example, if G is the graph of Fig. 1, $C=\{s, t\}$, edges of J are solid and edges of K are dotted, then $d_{P(G)}(x^J, x^K)=|J|=\min\{|J-K|, |K-J|\}$.

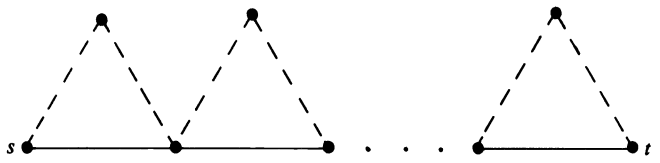


FIG. 1

COROLLARY 1. *The diameter of $P(G)$ is at most*

$$\max\{|J-K|: J, K \text{ are minimal postman sets}, |J-K|\leqq|K-J|\}.$$

Again, the graph of Fig. 1 shows that the bound of the corollary may be realized.

A d-dimensional polyhedron P with n facets has the *Hirsch property* if the diameter of P is at most $n-d$ (see [2, p. 168]). Klee and Walkup [5] have shown that unbounded polyhedra may not have the Hirsch property, but we now show that $P(G)$ does.

COROLLARY 2. *$P(G)$ has the Hirsch property.*

Proof. Since $\mathbb{R}_+^E$ is the recessional cone of $P(G)$, $P(G)$ has dimension $|E|$ and it is easily seen that for each $j\in E$, $x_j\geqq 0$ defines a "trivial" facet of $P(G)$. If J is a minimal postman set, then x^J is a vertex of $P(G)$, and it follows from elementary polyhedral theory that $P(G)$ must have at least $|J|$ nontrivial facets. The corollary now follows from Theorem 2. □

The postman polyhedron is closely related to the matching polyhedron. $J\subseteq E$ is a *perfect matching* if $|J\cap\delta(v)|=1$ for all $v\in V$. It is nontrivial to decide whether G has a perfect matching; Edmonds [3] gives an efficient algorithm for determining whether G has a perfect matching. Assuming that G has a perfect matching, the *matching polyhedron*, $M(G)$, is defined as conv $\{x^J: J$ is a perfect matching$\}$. Let $C=V$. (Since G has a perfect matching, $|C|$ is even). $M(G)$ is a face of $P(G)$, namely $M(G)=\{x\in P(G): \sum_{j\in E}x_j=|V|/2\}$. Hence, for perfect matchings J and K,

x^J and x^K are adjacent on $M(G)$

$\Leftrightarrow x^J$ and x^K are adjacent on $P(G)$

$\Leftrightarrow J\cup K$ contains the edge set of only one polygon

$\Leftrightarrow J\triangle K$ is the edge set of a connected subgraph of G.

This characterization of adjacency of vertices on $M(G)$ is also a special case of a result of Chvátal [1].

The components of the subgraph induced by $J\cup K$, for perfect matchings J and K, are even polygons and isolated edges. It follows from the above characterization of adjacency in $M(G)$ that $d_{M(G)}(x^J, x^K)\leqq|J\triangle K|/\alpha\leqq|V|/\alpha$, where α is the minimum length of an even polygon of G. Hence $M(G)$ has diameter at most $|V|/\alpha$ and it is a straightforward matter to construct graphs G such that $M(G)$ has diameter $|V|/\alpha$.

Suppose that G is a complete graph. Padberg and Rao [6] have shown that $M(G)$ has diameter 2. If $C = \{s, t\}$, then every minimal postman set is adjacent to the postman set consisting of the single edge $[s, t]$; so $P(G)$ has diameter 2. However, in general the postman polyhedron for a complete graph may have diameter greater than 2. Let G be the complete graph on node set $V = \{1, 2, \cdots, 7\}$, $C = \{1, 2, 3, 4\}$ and J and K be the minimal postman sets illustrated in Fig. 2, where edges of J are solid and those of K are dotted.

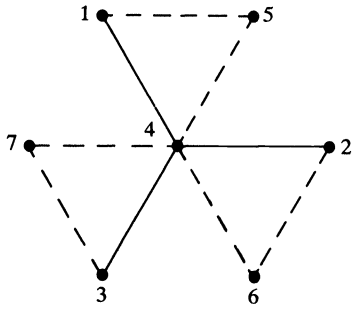


FIG. 2

If L is a third minimal postman set, then without loss of generality we can find edge sets A and B of 1–2 and 3–4 paths respectively such that $A \cap B = \emptyset$, $A \cup B = L$. From this it is a straightforward matter to check that $J \cup L$ or $K \cup L$ must contain the edge sets of at least two polygons. Therefore, by Theorem 1, L cannot be adjacent to both J and K, so $P(G)$ has diameter at least three.

REFERENCES

[1] V. CHVÁTAL, *On certain polytopes associated with graphs*, J.C.T. (B), 18 (1975), pp. 138–154.

[2] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.

[3] J. EDMONDS, *Paths, trees, and flowers*, Canad. J. Math., 17 (1965), pp. 449–467.

[4] J. EDMONDS AND E. L. JOHNSON, *Matching, Euler tours and the Chinese postman*, Math. Programming, 5 (1973), pp. 88–124.

[5] V. KLEE AND D. W. WALKUP, *The d-step conjecture for polyhedra of dimension d<6*, Acta. Math., 117 (1967), pp. 53–78.

[6] M. W. PADBERG AND M. R. RAO, *The travelling salesman problem and a class of polyhedra of diameter two*, Math. Programming, 7 (1974), pp. 32–45.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 2, June 1981

0196-5212/81/0202-0012 $01.00/0

SAMPLING SCHEMES FOR FOURIER TRANSFORM RECONSTRUCTION*

MARCI PERLSTADT†

Abstract. We study a problem that arises in radio astronomy and other fields. Astronomers measure a function in order to recover its Fourier transform. Restrictions on the locations of telescopes limit the number of measurements made. It is generally assumed that if the measurements obtained are in some sense "equispaced" then recovery of the transform will be satisfactory. We develop more objective criteria for evaluating the appropriateness of different sampling schemes. Applying these criteria to a simplified model we show that equispaced samples are not always optimal.

The problem can be abstracted as follows. Let f be a function and $\hat{f}$ its N-point finite Fourier transform. If we measure f at N appropriate points we can recover $\hat{f}$ using standard techniques. Suppose instead that f cannot be measured at N values. Assume one has some knowledge of $\hat{f}$, e.g., that f is band-limited so $\hat{f} = 0$ outside a band. Given L values of $\hat{f}$ we need only $p = N - L$ values of f to recover $\hat{f}$ in full. Now we ask: Which p values of f is it "best" to sample?

Our criterion for a sampling scheme to be good is that the computation of $\hat{f}$ be fairly insensitive to any sampling errors in f. We show that this is equivalent to studying the effect of perturbations on a matrix whose entries depend upon the values where f is sampled and the values where $\hat{f}$ is unknown. Different quantities associated with this matrix are studied to determine the effectiveness of the sampling scheme.

1. Introduction. We study a problem that arises in radio astronomy as well as in numerous other fields [1], [3], [5], [11]. Radio astronomers measure the visibility of a source in order to obtain its brightness distribution. It is known that brightness is the Fourier transform of visibility, so that if it were possible to measure the visibility at all points, the brightness could be recovered in a straightforward manner. In practice, the number of visibility samples is limited by the locations of radio telescopes. This is particularly significant for VLBI (Very Long Baseline Interferometry) studies, where sampling coverage is fairly sparse. It is generally assumed that if the samples obtained are in some sense "equispaced" then recovery of the brightness from the visibility is satisfactory. Our concern will be to develop criteria that will provide a more objective basis for determining the appropriateness of various sampling schemes. Using these criteria we show that equispaced samples are not always optimal.

The mathematical problem can be abstracted as follows. Let f be a function and $\hat{f}$ its N-point finite Fourier transform. If we could determine the value of f at N appropriate points then we could recover $\hat{f}$ using Fourier transform techniques. Suppose instead that we are unable to measure f at all N values but still wish to recover $\hat{f}$. Assume that one has some knowledge of $\hat{f}$, e.g., if f is band-limited then $\hat{f} = 0$ outside of a band. Given L values of f we need only measure $p = N - L$ values of f in order to recover the missing values of $\hat{f}$. The question arises as to which p values of f it would be "best" to sample. Our criterion for being a good sampling scheme is that the computation of $\hat{f}$ be fairly insensitive to any sampling errors in f.

It should be noted that there exist a large number of papers dealing with various aspects of the sampling problem for functions and their Fourier transforms. For a comprehensive review of these, see [7]. The particular question dealt with here is of a somewhat different nature.

Section 2 contains a formal statement of the problem. It is shown that the sensitivity of $\hat{f}$ to sampling errors in f can be measured by the condition number of a

* Received by the editors April 29, 1980. This work is based in part on the author's PhD thesis at the University of California at Berkeley.

† Department of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

matrix E whose entries depend upon the points where f is sampled and where $\hat{f}$ is unknown. In § 3 we show the conditions under which equispaced sampling is optimal. Clearly the quality of a sampling scheme depends upon the class of functions being sampled. For the general class of band-limited functions we show that equispacing is optimal. However, for appropriate subclasses of these functions we may have additional a priori knowledge of $\hat{f}$. Under these circumstances equispacing is frequently a poor scheme.

In § 4 we consider the case when we are able to sample more values of f than are needed in order to recover $\hat{f}$. This leads to an overdetermined system (i.e., E is rectangular) and we consider the problem from the point of view of least squares. We see that equispacing is no longer the only optimal sampling configuration for a band-limited function.

Section 5 deals with the problem of determining for which missing values of $\hat{f}$ it will be possible to sample f in such a manner that the resulting matrix E has a condition number of 1. It is shown that this is equivalent to looking at the roots of a certain polynomial with all its coefficients either 0 or 1.

Section 6 uses the material developed in § 5 to study a special case referred to as L gap L. By the L gap L case we mean the case where $\hat{f}$ is known outside of two bands of L points each. This is meant to be a simplified model of a double source. We show that by taking our samples in a certain nonequispaced fashion we can be assured that the condition number of E is less than or equal to $\sqrt{3}$.

Section 7 contains a brief discussion of alternative measurements of sensitivity to perturbations. Although the determinant of a matrix is generally a poor indicator of sensitivity, we show that for matrix E we can bound the condition number based on the determinant. This can save time in computer searches for small condition numbers.

Section 8 extends the results of earlier sections to the two-dimensional case. Once again, "equispaced" samples are shown to be optimal for "band-limited" functions, but not necessarily optimal in general. A two-dimensional extension of the L gap L case is also considered.

Section 9 contains a brief discussion of applications to radio astronomy. Given the widespread use of the finite Fourier transform throughout engineering, it is anticipated that many other applications exist.

2. Statement of the problem. Let f be a function of $\{0, 1, 2, \cdots, N-1\}$. Its discrete Fourier transform, $\hat{f}$ (also a function of $\{0, 1, 2, \cdots, N-1\}$), is given by

$$\hat{f}(k) = \frac{1}{N} \sum_{j=0}^{N-1} f(j)\omega^{-jk}, \tag{2.1}$$

where $\omega = e^{2\pi i/N}$. We know that

$$f(j) = \sum_{k=0}^{N-1} \hat{f}(k)\omega^{jk}, \tag{2.2}$$

and so $f(j)$ can be recovered from $\hat{f}(k)$. Note that (2.2) can be written in matrix form as

$$f = R\hat{f}, \tag{2.2}'$$

where f and $\hat{f}$ are column vectors with mth entry $f_m = f(m)$ and $\hat{f}_m = \hat{f}(m)$ and where R is $N \times N$ with mnth entry

$$R_{mn} = \omega^{(m-1)(n-1)}. \tag{2.3}$$

Suppose that $\hat{f}$ is known at l of the values $\{0, 1, 2, \cdots, N-1\}$ (as would be the case, for example, if f were known to be band-limited) and that we are to sample $f(j)$ at some points in order to recover the unknown values of $\hat{f}$. We need $p = N - l$ values of $f(j)$ to recover the rest of $\hat{f}$. The question dealt with here is: at which p values of j would it be "best" to sample f? By "best" we mean for which $j_1, j_2, \cdots, j_p$ will a "small" sampling error in $f(j_1), f(j_2), \cdots, f(j_p)$ result in only a "small" error in the computed values of $\hat{f}$. Let

$$(2.3)' \qquad \begin{aligned} f_1 &= (f(j_1), \cdots, f(j_p))^T, & f_2 &= (f(j_{p+1}), \cdots, f(j_N))^T, \\ \hat{f}_1 &= (\hat{f}(k_1), \cdots, \hat{f}(k_p))^T, & \hat{f}_2 &= (\hat{f}(k_{p+1}), \cdots, \hat{f}(k_N))^T, \end{aligned}$$

where $j_1, \cdots, j_p$ are the p values where f is sampled and $k_1, \cdots, k_p$ are the p values where $\hat{f}$ is not known. By rearranging the rows and columns of matrix R in (2.2)′ we obtain

$$(2.4) \qquad \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = \tilde{R} \begin{pmatrix} \hat{f}_1 \\ \hat{f}_2 \end{pmatrix} = \begin{pmatrix} E & F \\ G & H \end{pmatrix} \begin{pmatrix} \hat{f}_1 \\ \hat{f}_2 \end{pmatrix},$$

where $\tilde{R}$ is the $N \times N$ matrix with stth entry $\tilde{R} = \omega^{j_s k_t}$ and the blocks E, F, G, H are the result of partitioning $\tilde{R}$ so that E is $p \times p$, F is $p \times (N-p)$, G is $(N-p) \times p$ and H is $(N-p) \times (N-p)$. Thus

$$(2.5) \qquad E\hat{f}_1 = f_1 - F\hat{f}_2.$$

Since we assume that $\hat{f}_2$ was known exactly and that f_1 will be sampled, one measurement of the sensitivity of $\hat{f}_1$ to perturbations in f_1 is given by the condition number of E (see [4]).

Recall that the condition number of a $(p \times p)$ matrix A is $\mu(A) = \|A\|\|A^{-1}\|$. The norm we will use is the 2-norm. Thus $\mu(A) = \sqrt{\lambda p/\lambda_1}$, where $0 < \lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_p$ are the eigenvalues of A^*A (here $A^* = \bar{A}^T$, the conjugate transpose of A). The larger the quantity $\mu(A)$, the more sensitive the solution to the system $Ax = b$ is to perturbations in b [4]. It should be noted that having $\mu(A)$ large means that the problem itself (as opposed to a particular algorithm for solving the problem) is ill-conditioned. Thus if matrix E of (2.4) is ill-conditioned, then regardless of what method we use to solve for $\hat{f}_1$ (including exact arithmetic), the results obtained are still sensitive to perturbations in E and/or f_1.

From the above discussion we have that E is the $p \times p$ matrix with mnth entry $E_{mn} = \omega^{j_m k_n}$. Since the known frequencies are given, the unknown frequencies, $k_1, \cdots, k_p$, do not change. On the other hand, $j_1, j_2, \cdots, j_p$ correspond to the points where f is sampled, and we assume either that we are free to choose these or that we have at least some freedom in our choice of $j_1, \cdots, j_p$. Our goal is to find choices of $j_1, \cdots, j_p$ that make $\mu(E)$ as small as possible. As a first approximation we assume any choice of $j_1, \cdots, j_p$ is possible and try to minimize $\mu(E)$ for fixed $k_1, \cdots, k_p$.

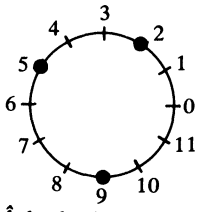
It is convenient to make our problem "continuous" by letting E have mnth entry $E_{mn} = \xi_m^n$ where $\xi_1, \cdots, \xi_p$ are points on the unit circle. Lemma 2.6 is immediate.

LEMMA 2.6. *For all $p \times p$ matrices of the form E, the trace of E^*E (denoted by* tr (E^*E)*) is p^2 and, therefore, $0 \leqq |E^*E| \leqq p^p$ (where $|E^*E|$ = determinant of E^*E). Thus $\mu(E) = 1$ if and only if $|E^*E| = p^p$ iff $E^*E = pI = EE^*$.*

Note that matrix R of (2.2)′ and (2.3) satisfies $\mu(E) = 1$.

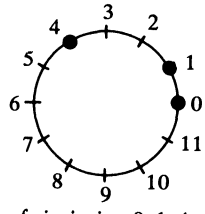
A convenient pictorial display for $j_1, \cdots, j_p$ and $k_1, \cdots, k_p$ (or, equivalently, $\xi_1, \cdots, \xi_p$ and $k_1, \cdots, k_p$) can be made using the unit circle. Associate with $j_1, \cdots, j_p$ the quantities $\omega^{j_1} = \xi_1, \cdots, \omega^{j_p} = \xi_p$, where $\omega = e^{2\pi i/N}$, and similarly associate

$k_1, \cdots, k_p$ with $\omega^{k_1}, \cdots, \omega^{k_p}$. Thus, for example, if $N = 12$, $j_1, j_2, j_3 = 0, 1, 4$ and $k_1, k_2, k_3 = 2, 5, 9$ we draw Figs. 1 and 2.



$\hat{f}$: $k_1, k_2, k_3 = 2, 5, 9$

FIG. 1



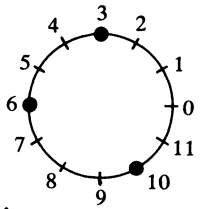
f: $j_1, j_2, j_3 = 0, 1, 4$

FIG. 2

With the aid of this pictorial representation we now state:

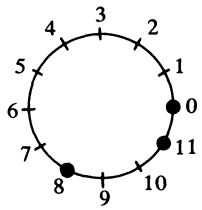
LEMMA 2.7. *Let P and P' be congruent polygons inscribed in the unit circle, with vertices $\xi_1, \cdots, \xi_p$ and $\xi'_1, \cdots, \xi'_p$ respectively. Let E and E' be $p \times p$ matrices with stth entries $E_{st} = (\xi_s)^{k_t}$ and $E'_{st} = (\xi'_s)^{k_t}$. Then $\mu(E) = \mu(E')$. In a similar fashion, if the polygon with vertices $\omega^{k_1}, \cdots, \omega^{k_p}$ is congruent to the polygon with vertices $\omega^{k'_1}, \cdots, \omega^{k'_p}$, then $k_1, \cdots, k_p$ in matrix E can be replaced by $k'_1, \cdots, k'_p$ without affecting $\mu(E)$.*

Example. The condition number of the matrix associated with Figs. 1, 2 is the same as the condition number associated with Figs. 3, 4.



$\hat{f}$: $k_1, k_2, k_3 = 3, 6, 10$

FIG. 3



f: $j_1, j_2, j_3 = 0, 8, 11$

FIG. 4

Proof. By an isometry of the unit circle we mean a one-to-one, distance-preserving map of the unit circle onto itself. Note that any isometry of the unit circle is completely determined by its action on any two points that do not lie on the same diameter. Thus every such isometry is either a rotation or a reflection about the x-axis possibly followed by a rotation. Note that $\mu(E)$ is unchanged by such a reflection or rotation of the ξ_i's. Since $\xi_1, \cdots, \xi_p$ and $\xi'_1, \cdots, \xi'_p$ are vertices of congruent p-gons if and only if there exists an isometry mapping $\{\xi_i\}_{i=1}^p$ onto $\{\xi'_i\}_{i=1}^p$, we are done. □

3. Equispacing. We return to the original question: Which p values of f do we sample? At first glance it may seem reasonable to choose p equispaced samples, i.e. choose $\xi_1, \cdots, \xi_p$ to be the pth roots of unity. In this section we discuss one case where this strategy works and then show that in most cases this is a very poor strategy.

A. *Band-limited functions.* Consider a function f where $\hat{f}$ is known outside of a band. This would be the case for the so-called "band-limited" functions for which $\hat{f}$ vanishes (and thus is known) outside, say, 0, 1, 2, $\cdots, p-1$. If p samples of f are measured at $\xi_1, \cdots, \xi_p$ then we have that E has stth entry

$$E_{mn} = \xi_m^{n-1}. \tag{3.1}$$

The following theorem tells us how to choose $\xi_1, \cdots, \xi_p$.

THEOREM 3.2. *For a function f with $\hat{f}$ known outside* $0, 1, 2, \cdots, p-1$, *the matrix E of* (3.1) *has a condition number of* 1 *if and only if the p samples of $f(\xi_1, \cdots, \xi_p)$ are equispaced.*

Proof. First assume that $\xi_1, \cdots, \xi_p$ are equispaced. By Lemma 2.7 we may assume $\xi_1 = 1, \xi_2 = \omega, \xi_3 = \omega^2, \cdots, \xi_p = \omega^{p-1}$, where $\omega = e^{2\pi i/p}$. Thus E has stth entry $E_{st} = \omega^{(s-1)(t-1)}$ which is exactly the form of matrix R in (2.3). Thus $\mu(E) = 1$.

Suppose now that E has the form of (3.1) and $\mu(E) = 1$. Thus $EE^* = pI$ by Lemma 2.6, and we may assume $\xi_1 = 1$ by Lemma 2.7. This means that the entries in the first column of E^* are just 1's, and since $EE^* = pI$ every row of E but the first must sum to 0; i.e., $1 + \xi_r + \cdots + \xi_r^{p-1} = 0$ for $r = 2, 3, \cdots, p$. Thus $\xi_2, \cdots, \xi_p$ satisfy $1 + x + x^2 + \cdots + x^{p-1} = 0$, and must be the pth roots of unity other than 1. □

An interesting consequence of this theorem is that it yields an alternative proof of the following well-known fact:

COROLLARY 3.3. *Suppose p points on the unit circle are to be chosen so as to maximize the product of their mutual distances. Then this product is maximized if and only if the p points are equispaced. The maximum value of the product is $\sqrt{p^p}$.*

Proof. Matrix E of (3.1) is Vandermonde, and so $|E^*E| = \prod_{1 \leq i < j \leq p} |\xi_i - \xi_j|^2$; i.e., $|E^*E|$ is the square of the product of all of the mutual distances points $\xi_1, \cdots, \xi_p$. Also, $0 \leq |E^*E| \leq p^p$ and $|E^*E| = p^p$ if and only if $\mu(E) = 1$. The result follows. □

B. *Arbitrary functions.* We now show that, in general, equispacing is not a good strategy. The basic result is stated below.

THEOREM 3.4. *Suppose $\hat{f}$ is unknown at $k_1, \cdots, k_p$, and p equispaced samples of f are measured at $\xi_1 = 1, \xi_2 = \omega, \xi_3 = \omega^2, \cdots, \xi_p = \omega^{p-1}$, where $\omega = e^{2\pi i/N}$. Then the corresponding matrix E has $|E| = 0$ (and thus $\mu(E) = \infty$) unless $k_1, \cdots, k_p$ are distinct* mod p. *If $k_1, \cdots, k_p$ are distinct* mod p *then $\mu(E) = 1$.*

Proof. E has stth entry $E_{st} = (\omega^{k_t})^{s-1}$. Thus E is Vandermonde and the magnitude of $|E|$ is given by $\prod_{1 \leq s < r \leq p} |\omega^{k_r - k_s}|$. If $k_1, \cdots, k_p$ are not distinct mod p, then $|E| = 0$. On the other hand, if $k_1, \cdots, k_p$ are distinct mod p, then E is just matrix R of (2.2)′ (up to some possible column interchanges) and so $\mu(E) = 1$. □

As an example consider the case when $N = 12$ and $\hat{f}$ is known outside of two bands of three points each. In particular let $k_1, \cdots, k_6 = 0, 1, 2, 6, 7, 8$ as shown in Fig. 5. Since 0, 1, 2, 6, 7, 8 are not distinct mod 6, by (3.4) equispaced sampling as in Fig. 6 yields $\mu(E) = \infty$. It will be shown later that sampling as in Fig. 7 yields $\mu(E) = 1$.

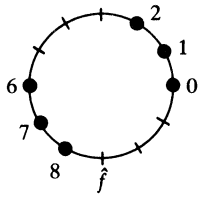


FIG. 5

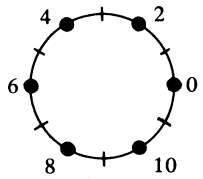


FIG. 6

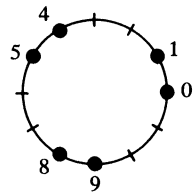


FIG. 7

4. Least squares and equispacing. We consider a modification of the situation described in §§ 2–3. Before, if l values of $\hat{f}$ were unknown then exactly l values of f were measured. In many applications this is an artificial constraint on the number of samples. In the following discussion we assume that at least l values of f are measured, thus introducing the possibility of an overdetermined system. A formal description follows.

Let $\hat{f}$ be unknown for $k_1, \cdots, k_l$ and be known exactly for $k_{l+1}, \cdots, k_N$. Suppose f is measured at $j_1, \cdots, j_p$ but not $j_{p+1}, \cdots, j_N$. Let

$$\begin{aligned} \hat{f}_1 &= (\hat{f}(k_1), \cdots, \hat{f}(k_l))^T, \qquad \hat{f}_2 = (\hat{f}(k_{l+1}), \cdots, \hat{f}(k_N))^T, \\ f_1 &= (f(j_1), \cdots, f(j_p))^T, \qquad f_2 = (f(j_{p+1}), \cdots, f(j_N))^T. \end{aligned} \tag{4.1}$$

Reordering the entries of matrix R of (2.2)′ and (2.3), we obtain (2.4), where $\tilde{R}_{mn} = \omega^{j_m k_n}$, $\omega = e^{2\pi i/N}$, and $\tilde{R}$ is partitioned into blocks E, F, G, H, where E is $p \times l$, F is $p \times (N-l)$, G is $(N-p) \times l$ and H is $(N-p) \times (N-l)$. Thus (2.5) holds.

To recover $\hat{f}_1$ we need N pieces of information. From f_1 and $\hat{f}_2$ we have $p+N-l$ pieces, so we assume $p \geqq l$. If $l = p$ we have exactly the situation described in § 2. If $p > l$, however, then we have more equations than unknowns in (2.5) and due to sampling errors in f_1 it is impossible to solve for $\hat{f}_1$ exactly. Instead one commonly requires that the solution to (2.5) be best in the "least squares" sense. Thus we choose $\hat{f}_1$ so as to minimize

$$\|E\hat{f}_1 - (f_1 - F\hat{f}_2)\|_2.$$

Once again we want to know the effect of perturbations in our measured values f_1, on the solution $\hat{f}_1$. One measurement of this sensitivity is given by the condition number of the rectangular matrix E (denoted by $\mu(E)$). See, for example, [6]. For the $p \times l$ matrix $E (p > l)$, $\mu(E) = \sqrt{\lambda_l/\lambda_1}$, where $0 < \lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_l$ are the eigenvalues of E^*E (assuming our underlying norm is the 2-norm). It should be noted that $\mu(E) = 1$ if and only if $E^*E = pI_l$, where I_l is the $l \times l$ identity matrix. It does not follow that if $E^*E = pI_l$ then $EE^* = lI_p$.

We can now return to the question of equispacing. We first consider the case when f is band-limited, i.e., $k_1, k_2, \cdots k_l = 0, 1, 2, \cdots, l-1$. We will see that if $\xi_1, \xi_2, \cdots, \xi_p$ are chosen equispaced, then $\mu(E) = 1$. On the other hand, this need not be the only choice of p points that makes $\mu(E) = 1$. This is in contrast to the situation of § 3 where $\mu(E) = 1$ if and only if the samples were equispaced. We have:

THEOREM 4.2. *Let f be band-limited, i.e., $k_1, \cdots, k_l = 0, 1, \cdots, l-1$. If f is sampled at $\xi_1, \cdots, \xi_p$ and E is the matrix with stth entry $E_{st} = \xi_s^{k_t} = \xi_s^{t-1}$, then $\mu(E) = 1$ if and only if $\xi_1, \cdots, \xi_p$ satisfy the $l-1$ equations*

$$\xi_1^i + \xi_2^i + \cdots + \xi_p^i = 0, \qquad i = 1, 2, \cdots, l-1. \tag{4.3}$$

Proof. The proof follows from the fact that $\mu(E) = 1$ if and only if $E^*E = pI_l$. □

If $p = l$, then note that the only solution to (4.3) (on the unit circle) is p equispaced points. This follows from Theorem 3.2, where we showed that when $p = l$, equispacing is the only scheme that makes $\mu(E) = 1$.

When $p > l$ many solutions exist. They will preserve some sense of "equispacing," since having $\xi_1 + \xi_2 + \cdots + \xi_p = 0$ means that $\xi_1, \xi_2, \cdots, \xi_p$ have center of gravity 0. As an example consider the case when $k_1, k_2 = 0, 1$ and $p = 4$. By Theorem 4.2, any choice of $\xi_1, \xi_2, \xi_3, \xi_4$ with center of gravity at 0 will make $\mu(E) = 1$. Thus, if $\xi_1, \xi_2, \xi_3, \xi_4$ are chosen as the endpoints of two diameters, $\mu(E) = 1$. For instance, let $\xi_1 = e^{i\theta_1}$, $\xi_2 = e^{i\theta_2}$, $\xi_3 = -\xi_1$, and $\xi_4 = -\xi_2$ (see Fig. 8).

In general, equispacing is not a good strategy for the least squares case. We have the following analogue of Theorem 3.4.

COROLLARY 4.4. *Let $\xi_1, \cdots, \xi_p = 1, \omega, \omega^2, \cdots, \omega^{p-1}$, where $\omega = e^{2\pi i/p}$. Let E be the $p \times l$ matrix with mnth element $E_{mn} = \xi_m^{k_n}$, where $k_1, \cdots, k_l$ are given. Then $\mu(E) = 1$ if and only if $k_1, \cdots, k_l$ are distinct* mod p. *If $k_1, \cdots, k_l$ are not distinct* mod p, *then $\mu(E) = \infty$.*

Proof. The proof is analogous to the proof of (3.4). □

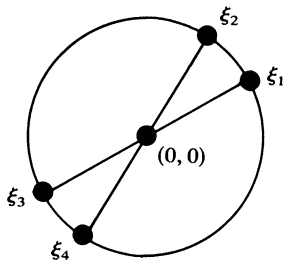


FIG. 8

5. When does $\mu(E) = 1$? In this section we examine the following problem: given $k_1, \cdots, k_p$, when do there exist $\xi_1, \xi_2, \cdots, \xi_p$ such that $\mu(E) = 1$? Although we cannot answer this question completely, the lemmas below provide criteria that are useful for specific cases. Furthermore, if we restrict to the discrete case ($\xi_j = e^{2\pi ik/N}$ for integers k and N) then there is an algorithm for deciding when $\mu(E)$ can be 1.

Two things should be noted. First, it is not always true that for a given $k_1, \cdots, k_p$ we can make $\mu(E) = 1$. For example, if $k_1, k_2, k_3 = 0, 1, 3$ then it can be shown [12] that $|E^*E| \leqq 20$ for all choices of ξ_1, ξ_2, ξ_3. Thus, by Lemma 2.6, $\mu(E) > 1$. Second, it is not necessary to have $\mu(E) = 1$, but only to have $\mu(E)$ relatively small. Our hope is that patterns of $\xi_1, \cdots, \xi_p$ that make $\mu(E) = 1$ for some $k_1, \cdots, k_p$ are similar to patterns of $\xi_1, \cdots, \xi_p$ that make $\mu(E)$ small for other $k_1, \cdots, k_p$ (see § 6).

Lemma 5.1 below is stated for the "discrete" case. Parts (i) and (ii) of the lemma generalize in the obvious way to the "continuous" case. The proof of Lemma 5.1 follows from Lemma 2.6.

LEMMA 5.1. *Let $k_1, \cdots, k_p$ be distinct nonnegative integers. Then* TFAE:

(i) *There exist N and $j_1, \cdots, j_p$ (distinct integers) so that every element of*

$$A = \{\omega^{j_m - j_n} | 1 \leqq n < m < p, \omega = e^{2\pi i/N}\}$$

satisfies the equation

$$x^{k_1} + x^{k_2} + \cdots + x^{k_p} = 0. \tag{5.2}$$

(ii) *There exist N and $j_1, \cdots, j_p$ (distinct integers) such that matrix E with mnth entry $E_{mn} = \omega^{j_m k_n} (\omega = e^{2\pi i/N})$ has $\mu(E) = 1$.*

(iii) *There exist N and $j_1, \cdots, j_p$ (distinct integers) such that each element of*

$$B = \{\omega^{k_m - k_n} | 1 \leqq m < n \leqq p, \omega = e^{2\pi i/N}\}$$

satisfies the equation

$$x^{j_1} + x^{j_2} + \cdots + x^{j_p} = 0.$$

Furthermore, N and $j_1, \cdots, j_p$ satisfy (i) *if and only if they satisfy* (ii) *if and only if they satisfy* (iii).

We note that, for fixed $k_1, \cdots, k_p$, there is an algorithm to test whether $\mu(E)$ can be 1 in Lemma 5.1. The existence of the algorithm depends upon several properties of cyclotomic polynomials [9]. We let $\Phi_N(x)$ be the Nth cyclotomic polynomial; i.e.,

$$\Phi_N(x) = \prod_{a \in A} (x - a), \quad \text{where } A = \{v^j | (j, N) = 1, v = e^{2\pi i/N}\}.$$

The degree of $\Phi_N(x)$ is $\phi(N)$, where ϕ is the Euler phi function. Note that, for fixed L, there exists M such that for all $M'>M$, $\phi(M')>L$. Thus there is an algorithm to determine all roots of unity which are also roots of (5.2). By Lemma 5.1 this yields the desired algorithm.

Several examples of the use of this lemma are given below.

Example 5.4. Let $k_1, \cdots, k_p = 0, 1, 2, \cdots, p-2, r$, where $0<p-2<r$. This is a natural generalization of the band-limited case, in that $\hat{f}$ is assumed to be known everywhere outside of a band except for the point r. It is of interest to see what happens as $k_p = r$ is moved about the circle. We have noted that for $k_1, k_2, k_3 = 0, 1, 3$, $\mu(E)$ is never 1. The following corollary can be proved using (5.1) [12].

COROLLARY 5.5. *Let* $k_1, \cdots, k_p = 0, 1, 2, \cdots, p-2, r$, *where* $0<p-2<r$. *Then unless* $p|r+1$, $\mu(E) \neq 1$. *Furthermore, if* $p|r+1$, *then* $\mu(E)=1$ *if and only if* $\xi_1, \cdots, \xi_p$ *are equispaced.*

Example 5.6. This example arises as a natural generalization of the band-limited case, and provides a very simplified one-dimensional model for a double source in radio astronomy. We assume $\hat{f}$ is known everywhere except in two bands of L points each. In terms of E, this corresponds to having $p=2L$ and $k_1, k_2, \cdots, k_{2L} = 0, 1, 2, \cdots, L-1, M, M+1, \cdots, M+L-1$, where $M \geqq L$. We refer to this case as L gap L. By (3.4), equispacing should be used if and only if $L \equiv M \pmod{2L}$. The question now becomes: what should be done if $L \not\equiv M \pmod{2L}$? We begin with

DEFINITION 5.7. Let $L_1, L_2, \cdots, L_r$ be r distinct regular L-gons inscribed in the unit circle. We call the set of vertices, $\xi_1, \cdots, \xi_p$ $(p=Lr)$, an *r-Lgon set.* Note that any r-Lgon set is completely determined by specifying one vertex from each of $L_1, \cdots, L_r$. This set of r vertices is called a *generator.*

Example. A 2-5gon set is shown in Fig. 9, where $\omega = e^{2\pi i/5}$. A generator is given by $\{1, \nu\}$. We assume from now on that the vertices $\xi_j = e^{i\theta_j}$, $j = 1, 2, \cdots, p$ are ordered so that $0 \leqq \theta_1 < \theta_2 < \cdots < \theta_p < 2\pi$. Thus $\{\xi_1, \cdots \xi_r\}$ is a generator.

COROLLARY 5.8. *For the case* L gap L $(k_1, \cdots, k_{2L} = 0, 1, \cdots, L-1, M, M+1, \cdots, M+L-1)$, *there exist* $\xi_1, \cdots, \xi_{2L}$ *such that the corresponding* E *satisfies* $\mu(E)=1$ *if and only if* $L|M$. *Furthermore, if* $L|M$ *then* $\mu(E)=1$ *if and only if* $\xi_1, \cdots, \xi_p$ *are isometric to the* 2-Lgon *set generated by* $\{1, \nu\}$, *where* ν *is an* Mth *root of* -1. □

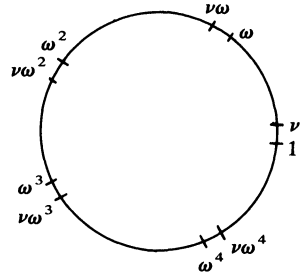


FIG. 9

Proof. The proof follows from Lemma 5.1. Details can be found in [12]. As an example we consider the case mentioned at the end of § 3, where $k_1, \cdots, k_6 = 0, 1, 2, 6, 7, 8$. Here $L=3$ and $M=6$. Thus $L|M$ and $\mu(E)=1$ if $\{\xi_1, \cdots, \xi_6\} = \{1, \omega, \omega^2, \nu, \nu\omega, \nu\omega^2 | \omega = e^{2\pi i/3}, \nu = e^{2\pi i/12}\}$ (see Figs. 5, 6, 7). In § 6 we show that for the

case L gap L we can have $\mu(E) \leqq \sqrt{3}$ (even if $L|M$) by choosing $\xi_1, \cdots, \xi_{2L}$ to be an appropriate 2-Lgon set.

It should be noted that the least squares case has its own analogue to Lemma 5.1.

6. A special case. We return to the L gap L case in Example 5.6. By Corollary 5.8, $\mu(E) = 1$ if and only if L/M and $\xi_1, \cdots, \xi_p$ form a certain 2-Lgon set. We now consider the case when $L|M$, and minimize $\mu(E)$ over all 2-Lgon sets. We find that this guarantees a way of choosing $\xi_1, \cdots, \xi_p$ so that $\mu(E) \leqq \sqrt{3}$. We have

COROLLARY 6.1. *Let* $k_1, \cdots, k_{2L} = 0, 1, \cdots, L-1, M, M+1, \cdots, M+L-1$, *where* $L|M$. *Then* $\mu(E)$ *is minimized over all* 2-L*gon sets by choosing* $\xi_1, \cdots, \xi_p$ *to be the* 2-L*gon set generated by* $\{1, \nu\}$, *where* $\nu = e^{2\pi i/(2Q-L)}$ *and* Q *is the (unique) integer such that* $Q \in \{M, M+1, \cdots, M+L-1\}$ *and* $L|Q$. *In this case* $[\mu(E)]^2 = (2+d)/(2-d)$, *where* $d = |1+\nu^Q|$ *and* $\mu(E) \leqq \sqrt{3}$.

The proof of Corollary 6.1 depends upon the fact that if $\xi_1, \cdots, \xi_p$ form an r-Lgon set ($p = Lr$) then E^*E (where $E_{mn} = \xi_m^{k_n}$) is block circulant with $r \times r$ blocks. Thus the eigenvalues of E^*E can be computed with relative ease. Details can be found in [12].

If ν is chosen as in Corollary 6.1, then the quantity $Q/(2Q-L)$ completely determines $\mu(E)$. As the gap between the two unknown bands of $\hat{f}$ grows (i.e., M increases), $Q/(2Q-L)$ gets "closer" to $\frac{1}{2}$ and $|1+\nu^Q|$ gets "closer" to 0. Thus as M increases the values of $[\mu(E)]^2$ decrease rapidly, as shown in Table 1.

TABLE 1

Q	$Q/(2Q-L)$	d	$[\mu(E)]^2$
2	2/3	1.	3.
3	3/5	.618034	1.894427
4	4/7	.445042	1.572417
5	5/9	.347296	1.420276
6	6/11	.284630	1.331858
7	7/13	.241073	1.274114

Thus, for example, given $k_1, \cdots, k_6 = 0, 1, 2, 10, 11, 12$ we have $L = 3$ and $Q = 12$. Thus $Q/(2Q-L) = \frac{4}{7}$, and if $\xi_1, \cdots, \xi_p$ are chosen as in Corollary 6.1 then $[\mu(E)]^2 \cong 1.572417$.

As an illustration of the relative sizes of $\mu(E)$ for the case L gap L, we computed the values of $\mu(E)$ when $k_1, \cdots, k_4 = 0, 1, 7, 8$ and $N = 48$. There 2,168 nonisometric cases (see § 2). Table 2 summarizes the results.

TABLE 2

Range of $[\mu(E)]^2$	Number of cases in this range
[10,000, ∞)	146
[1000, 10,000)	178
[500, 1000)	77
[100, 500)	386
[50, 100)	339
[25, 50)	396
[10, 25)	371
[5, 10)	192
[2, 5)	81
[1, 2)	2

7. Alternative measurements to $\mu(E)$. The condition number of E, $\mu(E)$ provides a "worst case" analysis of potential problems. It is often considered an overly pessimistic measure. A common alternative measurement of the sensitivity to sampling errors is provided by $K(E) = \text{tr}\ (E^*E)^{-1} = \sum_{k=1}^{p} 1/\lambda_k$, where $0 < \lambda_1, \cdots, \lambda_k$ are the eigenvalues of E^*E. It should be noted that, for the $p \times l$ matrix E, $K(E) \geqq l/p$ and $K(E) = l/p$ if and only if $\mu(E) = 1$.

Another quantity that is often used to measure the sensitivity of a system is the determinant of E^*E, $|E^*E|$ (here E is assumed to be $p \times p$). This, however, turns out to be a very unsatisfactory way of measuring this sensitivity. This is discussed, for example, in [4]. We will see below that in the case of matrix E, the quantity $|E^*E|$ can be of some use.

Recall that $|E^*E|$ attains its maximum value of p^p if and only if $\mu(E) = 1$. In fact, $|E^*E|$ can be used to provide an estimate for $\mu(E)$ when p is "small". This estimate in turn can be used to reduce computation time for test cases since $|E^*E|$ is somewhat cheaper to compute than $\mu(E)$. Again the fact that the eigenvalues, $0 \leqq \lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_p$, sum to p^2 provides the basis.

DEFINITION 7.1. The *determinant band* for a (given p) is the closed interval $[D_1, D_2]$, where D_1 is the minimum and D_2 the maximum of the quantity

$$\prod_{i=1}^{p} \mu_i,$$

subject to the constraints

$$\sum_{i=1}^{p} \mu_i = p^2, \quad \mu_p = a\mu_1, \quad 0 \leqq \mu_1 \leqq \mu_2 \leqq \cdots \leqq \mu_p,$$

where p is a given positive integer.

Thus the determinant band for a depends upon the choice of p. Let E be the usual $p \times p$ matrix with mnth entry $E_{mn} = \xi_m^{k_n}$. Note that if $[\mu(E)]^2 = a$ and $[D_1, D_2]$ is the determinant band for a (given p), then $D_1 \leqq |EE^*| \leqq D_2$.

LEMMA 7.2. *For fixed p, the determinant band for a is* $[D_1, D_2]$, *where*

$$D_1 = \min_{i=1,2,\cdots,p-1} \left\{ a^i \left[\frac{p^2}{ia + p - i} \right]^p \right\}$$

and

$$D_2 = \max \left\{ \frac{4ap^p}{(1+a)^2}, a^{p-1} \left(\frac{p^2}{pa - a + 1} \right)^p \right\}.$$

Proof. The proof follows from carrying out the maximization and minimization of Definition 7.1. Details can be found in [12]. □

COROLLARY 7.3. *Suppose* $1 \leqq a < a'$ *and p is fixed. Let* $[D_1, D_2]$ *be the determinant band for a and* $[E_1, E_2]$ *be the determinant band for a'. Then* $E_1 < D_1$ *and* $E_2 < D_2$.

Remark 7.4. Suppose that $|EE^*| = D$. Then we can solve the equation

$$D = \min_{i=1,\cdots,p-1} \left\{ a^i \left[\frac{p^2}{ia + p - i} \right]^p \right\} \tag{7.5}$$

for a and call the solution C_1. Note that C_1 yields the smallest possible condition number for matrix E. We can also solve the equation

$$D = \max \left\{ \frac{4ap^p}{(1+a)^2}, a^{p-1} \left(\frac{p^2}{pa - a + 1} \right)^p \right\} \tag{7.6}$$

for a and call the solution C_2. C_2 is the largest condition number possible when $|EE^*| = D$. We can call $[C_1, C_2]$ the "condition number band" for D. Thus if $|EE^*| = D$ then $C_1 \leqq \mu(E) \leqq C_2$. It should be noted that in practice (i.e., numerically) (7.5) and (7.6) are easy to solve since the right-hand side of each is a strictly decreasing function of a for $a > 1$.

Example. Let $K_1, K_2, K_3, K_4 = 0, 1, 5, 6$. Suppose we wish to search for the smallest $\mu(E)$ over all choices of J_1, J_2, J_3, J_4, where $N = 48$. There are a total of 2,168 nonisometric cases to consider. If we were to use $N = 24$ instead of $N = 48$ there would be only 256 nonisometric cases to consider. Suppose we first run the 256 cases for $N = 24$, computing only $|EE^*|$. We find $|EE^*| \leqq 195.98$. Using Lemma 7.2, we get that $D \leqq 195.24$ if and only if $a \geqq 2.9$. We can now turn our attention to the 2,168 cases arising when $N = 48$. Instead of computing $\mu(E)$ for each case, we will compute $|EE^*|$. Since the determinant band for 2.9 is roughly [148.9009, 195.24], we only compute $\mu(E)$ if $|EE^*| \geqq 148.9$. The result is that while we compute $|EE^*|$ for all 2,168 cases we need only compute $\mu(E)$ for 60 cases. This yields a reduction in computation time.

8. The two-dimensional case. In this section we consider the two-dimensional analogue of the problem discussed in § 2. The situation becomes much more complicated, and the results obtained are not as strong as those obtained for the one-dimensional case. We proceed with a formal statement of the two-dimensional problem.

Let f be a function on $\{0, 1, 2, \cdots, N-1\} \times \{0, 1, 2, \cdots, N-1\}$. Its discrete Fourier transform, $\hat{f}$ (also a function of $\{0, 1, \cdots, N-1\} \times \{0, 1, \cdots, N-1\}$), is given by

$$\hat{f}(m, n) = \frac{1}{N^2} \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} f(j, k) \omega^{-(jm+kn)}, \tag{8.1}$$

where $\omega = e^{2\pi i/N}$. In turn

$$f(j, k) = \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} \hat{f}(m, n) \omega^{(jm+kn)}, \tag{8.2}$$

so that $f(j, k)$ can be recovered from $\hat{f}(m, n)$. Just as in § 1, (8.2) can be rewritten in matrix form as $f = R\hat{f}$, where the entries of f and $\hat{f}$ are in lexicographic order and $R_{jk} = \omega^{v_j \cdot v_k}$, where v_i is the ith 2-tuple in the lexicographic ordering of $\{0, 1, 2, \cdots, N-1\} \times \{0, 1, 2, \cdots, N-1\}$ and $\omega = e^{2\pi i/N}$.

We consider the two-dimensional analogue of the original problem. Namely, we assume $\hat{f}$ is known at all but p points, and we sample f at p points. Let

$$f_1 = (f(j_1, k_1), \cdots, f(j_p, k_p))^T, \qquad f_2 = (f(j_{p+1}, k_{p+1}), \cdots, f(j_{N^2}, k_{N^2}))^T,$$
$$\hat{f}_1 = (\hat{f}(m_1, n_1), \cdots, \hat{f}(m_p, n_p))^T, \quad \hat{f}_2 = (\hat{f}(m_{p+1}, n_{p+1}), \cdots, \hat{f}(m_{N^2}, n_{N^2}))^T,$$

where $(j_1, k_1), \cdots, (j_p, k_p)$ are the points where f is sampled and $(m_1, n_1), \cdots, (m_p, n_p)$ are the points where $\hat{f}$ is unknown. By rearranging the rows and columns of matrix $\tilde{R}$ we obtain (2.4), where R is the $N^2 \times N^2$ matrix with stth entry $\tilde{R}_{st} = \omega^{j_s m_t + k_s n_t}$ and E, F, G, H are blocks of the appropriate size (E is $p \times p$). Thus we have (2.5).

Once again we want to study the effect of perturbations in the sampled values f_1 by considering $\mu(E)$. Note that we still have $\mu(E) = 1$ if and only if $E^*E = pI$. As in the one-dimensional case we may make the problem continuous by letting $\omega^{j_s} = \xi_s$ and $\omega^{k_s} = \nu_s$. Thus E has stth entry $E_{st} = \xi_s^{m_t} \nu_s^{n_t}$, where $\xi_s, \nu_s \in D = \{e^{i\theta} | 0 \leqq \theta < 2\pi\}$. Thus the problem becomes for fixed $(M_1, N_1), \cdots, (M_p, N_p)$, choose $(\xi_1, \nu_1), \cdots, (\xi_p, \nu_p) \in D \times D$ so that $\mu(E)$ is as small as possible.

DEFINITION 8.3. The set $\{(j_{r'}, k_{r'})|r = 1, \cdots, p\}$ is a translation (mod N through (J, K)) of the set $\{(j_r, k_r)|r = 1, \cdots, p\}$ if $j_{r'} = j_r + J \pmod N$ and $k_{r'} = k_r + K \pmod N$ for $r = 1, \cdots, p$.

Note that $\mu(E)$ is unchanged by translations of $\{(j_r, k_r)|r = 1, \cdots, p\}$ or $\{m_r, n_r|r = 1, \cdots, p\}$. For the "continuous" problem this is equivalent to saying that $\mu(E)$ is unchanged if $\{(\xi_r, \nu_r)|r = 1, \cdots, p\}$ is replaced by $\{\alpha\xi_r, \beta\nu_r|r = 1, \cdots, p$ and $\alpha, \beta \in D\}$. Thus we may assume $(\xi_1, \nu_1) = (1, 1)$.

It is now logical to ask how well equispacing works in the two-dimensional case. In this case it is not as clear what is meant by equispaced samples, and the problem is a bit more complicated. Definition 8.3 will provide us with an intuitively reasonable way of choosing L^2 equispaced points. In general, however, it is not clear what the "right" way to equispace p points is, when p is not an integer squared. For this reason the various examples that are worked below were carefully chosen so as to have an obvious way of picking an equispaced set. Consider the following:

DEFINITION 8.4. Let $R = \{1, \rho, \rho^2, \cdots, \rho^{L-1}|\rho = e^{2\pi i/L}\}$. We call the L^2 points of the set $R \times R$ (as well as any translation of these L^2 points) a set of L^2 *equispaced points.*

It is clear that the L^2 equispaced points defined above agree with our intuitive notion of how L^2 equispaced points should be placed. As an example, consider the two 36 "point" grids (Figs. 10, 11) (where each grid point is represented by a square). The darkened squares represent sets of 9 equispaced points.

We will also wish to consider sets of what we will call nearly equispaced points.

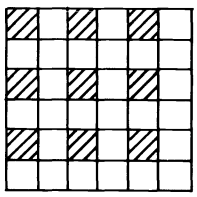


FIG. 10

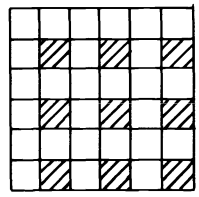


FIG. 11

DEFINITION 8.5. A set of L^2 points is called *nearly equispaced* if it is of the form

$$\{(\alpha_l\rho^j, \rho^l)|\alpha_0, \cdots, \alpha_{L-1} \in D, 0 \leqq j \leqq L-1, \rho = e^{2\pi i/L}\}, \tag{8.5a}$$

or of the form

$$\{\rho^l, \beta_l\rho^k|\beta_0, \cdots, \beta_{L-1} \in D, 0 \leqq k \leqq L-1, \rho = e^{2\pi i/L}\} \tag{8.5b}$$

or is produced by a translation of (8.5a) or (8.5b). Recall that $D = \{e^{i\theta}|0 \leqq \theta < 2\pi\}$.

Thus the 9 darkened squares on the two 6×6 grids (Figs. 12, 13) represent two 9-point nearly equispaced sets.

We now consider what happens when f is "band-limited". For our purpose it is convenient to define band-limited as follows:

DEFINITION 8.6. A function f is *band-limited* if $\hat{f}$ vanishes outside of an $L \times L$ square. Without loss of generality, we may assume $\hat{f}$ vanishes outside the $L \times L$ square $\{(j, k)|0 \leqq j, k \leqq L-1\}$.

THEOREM 8.7. *Let f be band-limited, so that $\hat{f}$ vanishes outside of the $L \times L$ square of* (8.6). *Suppose f is sampled at $(\mu_1, \nu_1), \cdots, (\mu_{L^2}, \nu_{L^2})$ (here $\mu_i, \nu_i \in D = \{e^{i\theta}|0 \leqq 2\pi\}$) and the corresponding $L^2 \times L^2$ matrix E is formed. Then $\mu(E) = 1$ if and only if $\{(\mu_i, \nu_i)|i = 1, \cdots, L^2\}$ is equispaced or nearly equispaced.*

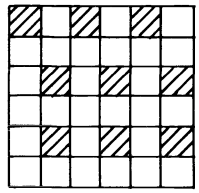


FIG. 12

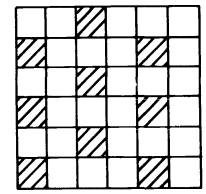


FIG. 13

Proof. The proof follows from the fact that $\mu(E)=1$ *if and only if* $E^*E=pI$. Details can be found in [12]. □

In general, "equispaced" configurations need not be optimal. We can see this by considering a two-dimensional analogue of the L gap L case discussed in §§ 5 and 6. Assume that $\hat{f}$ is known outside to two nonintersecting squares, each of size $L\times L$. Thus without loss of generality we have

$$\{(M_i, N_i)|i=1,2,\cdots,2L^2\}=(E_1\times E_1)\cup(E_2\times E_3),$$

where $E_1=\{0,1,2,\cdots,L-1\}$, $E_2=\{R,R+1,\cdots,R+L-1\}$ and $E_3=\{S,S+1,\cdots,S+L-1\}$. This case will be referred to as $\boxed{L}$ gap $\boxed{L}_{RS}$ (the RS subscript indicating the corner where the second square begins) or just $\boxed{L}$ gap $\boxed{L}$. For example, in Fig. 14 the darkened squares represent $(E_1\times E_1)\cup(E_2\times E_3)$ for the case $\boxed{2}$ gap $\boxed{2}_{5,6}$.

We would like to know how to choose $\{(\xi_i, \nu_i)|i=1,2,\cdots,2L^2\}$ so that $\mu(E)$ is small. We will see in the example below that an "equispaced" choice can be a very poor solution.

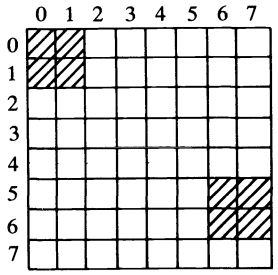


FIG. 14

Suppose $R=S=3$ for the case $\boxed{3}$ gap $\boxed{3}$ described above. On a 36-point grid this is represented by the darkened squares of Fig. 15. Let $\{(\xi_i, \nu_i)|i=1,2,\cdots,2L^2\}=(F_1\times F_1)\cup(F_2\times F_2)$, where $F_1=\{1,\omega^2,\omega^4\}$ and $F_2=\{\omega,\omega^3,\omega^5\}(\omega=e^{2\pi i/6})$. This choice of (ξ_i, ω_i) agrees with our intuitive notion of "equispacing", yielding a checkerboard pattern on the 36-point grid in Fig. 16. The resulting matrix E is singular and so $\mu(E)=\infty$.

From this example we can see that there is no reason to assume, a priori, that "equispacing" is a good strategy. We have the following analogue of Corollary 5.8.

LEMMA 8.8. *For the case* $\boxed{L}$ gap $\boxed{L}_{R,S}$, $\mu(E)=1$ *if and only if* $\{(\xi_i, \nu_i)|i=1,2,\cdots,2L^2\}$ *is chosen so that for* $i\neq j$, $(\xi_i\bar{\xi}_j, \nu_i\bar{\nu}_j)$ *is a root of*

$$q(x,y)=(1+x+x^2+\cdots+x^{L-1})(1+y+y^2+\cdots+y^{L-1})(1+x^ry^s)=0.$$

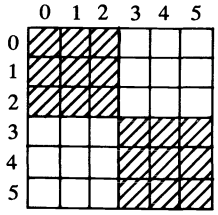


FIG. 15

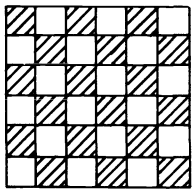


FIG. 16

Suppose that $L|R$ and $L|S$ and we choose $(\xi_1, \nu_1), \cdots, (\xi_{L^2}, \nu_{L^2})$ to be an equispaced set of L^2 points ($\xi_1 = \nu_1 = 1$) and $(\xi_{L^2+1}, \nu_{L^2+1}), \cdots, (\xi_{2L^2}, \nu_{2L^2})$ to be a second equispaced set of L^2 points, where $(\xi_{L^2+1}, \nu_{L^2+1})$ satisfies $q(x, y) = 1 + x^r y^s = 0$. Then it is easy to see that $\mu(E) = 1$. It should be noted that this choice of (ξ_i, ν_i) is very much analogous to the 2-Lgon sets of Definition 5.7. Recall that for the case L gap L it was 2-Lgon sets that made it possible to have $\mu(E) = 1$.

If we return to the case $\boxed{L}$ gap $\boxed{L}_{3,3}$ discussed earlier, we note that by Lemma 8.8 we have that if $\{(\xi_i, \nu_i) | i = 1, 2, \cdots, 2L^2\} = (F_1 \cup F_2) \times F_1$, then $\mu(E) = 1$. For a 6×6 grid this yields the following sampling pattern (see Fig. 17).

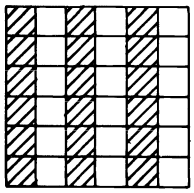


FIG. 17

It should be noted that the two-dimensional case can be generalized in the obvious way to the "least squares" problem discussed in § 4. Although we have been unable to state anything as strong as Theorem 3.4 or Corollary 4.4 for the two-dimensional case, the examples presented do show that it cannot be assumed that equispaced samples are optimal or even satisfactory.

9. Applications. The Fourier transform arises in many contexts in engineering and science. See, for example, [1], [3], [5] and [11]. We will discuss a particular application to radio astronomy.

Radio astronomers measure the visibility of a source, $V(u, v)$, in order to obtain its brightness distribution, $I(x, y)$. The quantities $V(u, v)$ and $I(x, y)$ are related by the equation

$$I(x, y) = \iint V(u, v) \exp\{-2\pi i(ux + vy)\}\, du\, dv, \tag{9.1}$$

and so $I = \hat{V}$ [3]. If one were able to measure $V(u, v)$ at all points, then $I(x, y)$ could be recovered from $V(u, v)$ in a straightforward manner. In practice the number of visibility samples is limited by the locations of radio telescopes. This is particularly significant for VLBI (Very Long Baseline Interferometry) studies, where sampling coverage is very sparse. These studies are done by using radio telescopes at various locations across the

continent (hence the very long baseline). The results are increased resolution and sparse sampling coverage of the u, v-plane.

The values of the argument (u, v) where V is actually measured are determined by the relative locations of the telescopes and the location of the source. For every pair of telescopes, T_2 and T_1, let

B = distance from T_2 to T_1,

D = angle of declination of the vector from T_2 to T_1, $\overrightarrow{T_2T_1}$,

δ = angle of declination of the vector from the center of the earth to the source.

By the angle of declination of the vector $\vec{ab}$ we mean the following: Translate vector $\vec{ab}$ so that point a lies at the center of the earth. The angle that $\vec{ab}$ makes above or below the equatorial plane is the angle of declination.

In the course of 24 hours, the pair of telescopes, T_2 and T_1 traces out measurements of V along the path of the ellipse given by

$$\frac{u^2}{a^2}+\frac{(v-v_0)^2}{b^2}=1, \tag{9.2}$$

where $a = B \cos D$, $n = B \cos D \sin \delta$, and $v_0 = B \sin D \cos \delta$. For a derivation of these equations see [3]. Every pair of telescopes traces out an ellipse of this form, and this determines our sampling coverage.

Once the samples are obtained, a variety of procedures are used to determine $I(x, y)$ in an attempt to compensate for the fairly sparse coverage of $V(u, v)$. These are discussed, for example, in [3] and [8]. In the end one obtains an approximation $I'(x, y)$ for $I(x, y)$.

Regardless of which procedure is used, the question arises as to how well the function $I'(x, y)$ that is determined in the end actually approximates $I(x, y)$. In general there are some aspects of $I(x, y)$ about which the astronomer is quite certain and other aspects where there is much uncertainty. For instance, it may be clear that what is involved is a small diameter source, in which case $I(x, y)$ is band-limited, or that we have a double source as in the $\boxed{L}$ gap $\boxed{L}$ case. If we wish to know how sensitive our final answer $I'(x, y)$ is to errors made in sampling $V(u, v)$, we could proceed as follows. Consider the points where the astronomer is fairly certain of $I(x, y)$. Call these points $k_{l+1}, k_{l+2}, \cdots, k_{N^2}$, where $k_1 \in \mathbb{R}^2$, $i = 1, 2, \cdots, N^2$. We assume the points where V was actually sampled are labeled $j_1, j_2, \cdots, j_p$, where $j_i \in \mathbb{R}^2$, $i = 1, 2, \cdots, N^2$. We now consider how sensitive the computation of the remaining l values of $I(x, y)$ is to error in the sampled values of V. This is exactly the situation described in § 8.

Another problem that arises deals with the location of a new telescope to be built in the Midwest for use in VLBI studies. A number of reports [2], [8], [14] have dealt with this problem. The reports show u, v-traces for existing radio telescopes for sources at several declinations and then show the additional u, v-traces obtained if the new telescope is placed in various locations. The best location is deemed to be the one that provides the most "equispaced" coverage of the u, v-plane. Although the model discussed in § 8 is greatly simplified and provides only a first approximation to the actual situation, it indicates that in certain cases there may be little objective basis to seeking the most "equispaced" coverage.

Note that when one is using the model studied in the last two sections, the adequacy of sampling coverage would vary with the nature of the source (e.g., double source versus small diameter source), as it is the nature of the source that determines a priori knowledge of $I(x, y)$. Thus (not surprisingly) it would be possible for a scheme that works well for one source to work poorly for another source.

Acknowledgments. This paper is based upon results of a doctoral thesis. It is a pleasure to express my deep appreciation to my advisor, Professor F. Alberto Grünbaum, for his suggestions, guidance and advice.

REFERENCES

1. R. BRACEWELL, *The Fourier Transform and Its Applications*, McGraw-Hill, New York, 1965.
2. M. M. COHEN, ed. *VLBI Network Studies, Vol. I: A VLBI Network Using Existing Telescopes*, California Institute of Technology, Owens Valley Radio Observatory, 1975.
3. E. B. FOMALONT AND M. C. WRIGHT, *Interferometry and aperture synthesis*, in Galactic and Extra Galactic Radio Astronomy, Springer-Verlag, New York, 1974.
4. G. FORSYTHE AND C. B. MOLER, *Computation of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1967.
5. J. W. GOODMAN, *Introduction to Fourier Optics*, McGraw-Hill, San Francisco, 1968.
6. R. J. HANSON AND C. L. LAWSON, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974.
7. A. J. JERRI, *The Shannon sampling theorem–its various extensions and applications: a tutorial review*, Proc. IEEE, 65 (1977), pp. 1565–1596.
8. K. I. KELLERMAN, ed., *VLBI Network Studies, Vol. III: A Dedicated VLBI Network*, National Radio Astronomy Observatory, Green Bank, West Virginia, 1977.
9. S. LANG, *Algebra* (1971), Addison-Wesley, Reading, MA, 1971.
10. T. MUIR, *A Treatise on the Theory of Determinants*, Longmans, Green, New York, 1933.
11. A. V. OPPENHEIM AND R. W. SCHAFER, *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975.
12. M. PERLSTADT, *A Study of Sampling Schemes for Fourier Transform Reconstruction with an Error Analysis*, PhD thesis, University of California, Berkeley, CA, 1978.
13. G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.
14. G. W. SWENSON ET AL. eds., *VLBI Network Studies, Vol. II: Interim Report on a New Antenna for the VLBI Network*, University of Illinois at Urbana, Vermilion River Observatory, 1977.
15. J. W. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 2, June 1981

0196-5212/81/0202-0013 $01.00/0

HALF-PLANE MINIMIZATION OF MATRIX-VALUED QUADRATIC FUNCTIONALS*

PH. DELSARTE,† Y. GENIN†‡ AND Y. KAMP†

Abstract. The general subject of the paper is the minimization over space functions with half-plane supports of certain quadratic functionals defined from a two-variable Hermitian-valued measure. A detailed theory of the minimizing functions is developed for the particular situation where the supports are intervals in the lexicographic ordering of the integer plane. The main results are concerned with topics such as stability properties, recurrence relations and spectral factorization, which play a significant role in digital signal processing and estimation theory.

1. Introduction and notation. This paper contains a generalization to matrix-valued functions of the theory of half-plane Toeplitz systems [6]. It can also be viewed as a two-variable extension of the theory of matrix polynomials orthogonal on the unit circle [4], [21]. Emphasis is put on properties of recursive stability and computability, which are very important for applications in digital signal processing and estimation theory [8], [17], [20]. In this context significant topics are spectral factorization and its approximations [3], [8], [16].

The scope of the paper is restricted to half-plane systems, which as shown by Helson and Lowdenslager [11] constitute the simplest generalization of the classical one-variable systems. In that respect, let us especially mention Hirschman's contribution [12], [13], containing an appropriate extension of Szegö's orthogonal polynomials to the half-plane situation (as part of a very general theory). In a recent paper, Marzetta [16] studied these "half-plane orthogonal polynomials" from a different point of view and emphasized their great significance regarding applications in linear prediction theory. The present paper borrows several ideas from Marzetta's approach, dealing with two-variable functions of lexicographic interval support that minimize a given quadratic functional. It turns out that the main properties of the matrix-valued minimizing functions with such supports can be easily deduced from certain results belonging to the theory of block-Hankel operators developed by Adamjan, Arov and Krein [1]. The necessary material is given in a recent paper by the authors [7]. Most results herein are obtained in the framework of the Hilbert–Lebesgue space L_2, which appears to be naturally adapted to our approach. However, the theory can also be successfully developed in the more restrictive but very interesting context of the Banach algebras first considered by Baxter [2] in the scalar case and recently extended to the matrix case by Geronimo [9]. (See also Hirschman [12]–[15].)

After having mentioned the general background of the present work we will describe its organization and emphasize some significant results. (These are obtained under ad hoc technical assumptions which are not specified in this Introduction.)

Section 2 introduces the main theme of the paper, namely minimization in prescribed function spaces of two quadratic functionals defined in terms of a given Hermitian-valued measure. The existence, uniqueness and characterization of the minimizing functions are established in the simple case where the space support is a finite subset of the integer plane.

* Received by the editors October 29, 1980.

† Philips Research Laboratory, Av. Van Becelaere 2, B-1170, Brussels, Belgium.

‡ The work of this author was supported in part by the U.S. Army Research Office under contract DAAG 29-79-C-0215, the Joint Services Electronics Program under contract DAAG 29-79-C-0047 and the National Science Foundation under grant NSF-ENG 78-1003, while the author was on leave at Information Systems Laboratory, Stanford University, Stanford, California.

Section 3 contains preliminary material from another source, namely the minimization of one-variable Hermitian functionals depending on a parameter. The normalized minimizing polynomials are defined, via spectral factorization, and their main algebraic properties are reported: orthogonality and recurrence relations, partial moment reconstruction.

In § 4 the general half-plane minimization problem is investigated in the situations where the space support is allowed to be infinite but restricted to have finite width. The results quoted in § 2 are shown to remain valid in such cases. In particular, when the support is a horizontal strip the minimizing functions are identified to be the matrix polynomials of § 3; this fact plays an important role in the general theory. The last result is concerned with convergence in the L_2-norm of sequences of minimizing functions with increasing or decreasing supports.

Section 5 is central. It contains a thorough study of those minimizing functions the supports of which are intervals in the lexicographic ordering of the integer plane. In § 5.1 it is first shown how these particular minimizing functions can be expressed in terms of the parametric polynomials of § 3 by means of a well-defined J-unitary matrix function of the parameter variable. Section 5.2 is devoted to the derivation of the stability properties of the minimizing functions, which leads to a partial reconstruction of the measure. Section 5.3 is concerned with the important Schur parameters occurring in the three-term recurrence relations connecting the minimizing functions of interval support. The stability properties are shown to be hereditary with respect to these relations, for any values of the Schur parameters.

The whole theory is reexamined in § 6 under certain summability assumptions regarding either the measure function or the Schur parameters. Sequences of minimizing functions of interval support tending to a horizontal strip are proved to be convergent in a strong sense. Next, construction of the theory from the Schur parameters is described. Finally it is shown how the interval functions can be approximated by asymptotically stable minimizing trigonometric polynomials.

Although it is implicitly present throughout the paper, the question of half-plane spectral factorization is treated as such only in § 7. It is pointed out that the canonical spectral factor of the derivative of the measure function can be inversely approximated by any sequence of minimizing trigonometric polynomials the supports of which tend to the whole half-plane. This allows one to obtain the inverse of this spectral factor as the limit of some interesting families of asymptotically stable minimizing functions.

Notation

$A_j(e^{i\theta}, w)$, $B_j(e^{i\theta}, w)$	parametric minimizing polynomials
$M_j(e^{i\theta})$, $N_j(e^{i\theta})$	normalizing factors for A_j, B_j
$\Omega_j(\theta)$	contractive matrix function
$\Gamma_j(\theta)$	block–Toeplitz matrix function
J	diagonal matrix $I \dotplus (-I)$
$U_j(e^{i\theta})$, $U_{m,j}(e^{i\theta})$, $Z_{m,j}(e^{i\theta})$	J-unitary matrix functions
$F\ldots(e^{i\theta}, w)$, $G\ldots(e^{i\theta}, w)$	minimizing functions of support $H\ldots$
$X\ldots(e^{i\theta}, w)$, $Y\ldots(e^{i\theta}, w)$	normalized minimizing functions
H	upper half-plane in $\mathbb{Z}\times\mathbb{Z}$
H_j	horizontal strip of width j in H
$H_{m,j}$	inverse lexicographic interval in H
$H_{k,m,j}$	finite subset of $H_{m,j}$ (defined by $\lvert s\rvert \leqq k$)
$E_{m,j}$	Schur parameter (= contractive matrix)

2. Minimization for finite support. Let Σ be a Hermitian-valued measure in two variables. By definition, $\Sigma(\theta, \phi)$ is a $p \times p$ Hermitian matrix function, defined on the square $0 \leqq \theta, \phi \leqq 2\pi$, satisfying the nondecrease condition $\Sigma(\theta_1, \phi_1) \leqq \Sigma(\theta_2, \phi_2)$ when $\theta_1 \leqq \theta_2$ and $\phi_1 \leqq \phi_2$. (Throughout this paper the notation $A \leqq B$ is used to mean that $B - A$ is nonnegative definite.) With the measure Σ we associate both right and left matrix-valued inner products

$$(1)\qquad \begin{aligned}(P, Q)_r &= \frac{1}{4\pi^2}\int\int_0^{2\pi} \tilde{P}(e^{i\theta}, e^{i\phi})\, d\Sigma(\theta, \phi) Q(e^{i\theta}, e^{i\phi}),\\ (P, Q)_l &= \frac{1}{4\pi^2}\int\int_0^{2\pi} Q(e^{i\theta}, e^{i\phi})\, d\Sigma(\theta, \phi) \tilde{P}(e^{i\theta}, e^{i\phi}),\end{aligned}$$

where P and Q vary over a class of $p \times p$ matrix-valued functions with suitable properties. (For a thorough treatment of this subject the reader is especially referred to Rosenberg [18].) Note that $(P, P)_r \geqq 0$ and $(P, P)_l \geqq 0$ for all P.

Next let us define the right and left Hermitian-valued functionals Φ_r and Φ_l as follows:

$$(2)\qquad \Phi_{r,l}(P) = (P, P)_{r,l} - 2 \operatorname{Herm} P(0, 0),$$

where $P(0, 0) = (4\pi^2)^{-1} \iint P(e^{i\theta}, e^{i\phi})\, d\theta\, d\phi$ is the mean value of $P(e^{i\theta}, e^{i\phi})$ and $\operatorname{Herm} A = (A + \tilde{A})/2$ denotes the Hermitian part. The subscripts r and l appearing in (1) and (2) will often be dropped in the sequel. We are interested in the problem of minimizing the functional Φ over a given space L of functions. Specifically, a given $P \in L$ is said to *minimize* Φ over L if it satisfies $\Phi(P) \leqq \Phi(Q)$ for all $Q \in L$.

Let S be a finite subset of the integer plane $\mathbb{Z} \times \mathbb{Z}$. Assuming S to contain the origin $(0, 0)$, let us write $S = \{(s_0, t_0), (s_1, t_1), \cdots, (s_n, t_n)\}$ with $s_0 = t_0 = 0$. A trigonometric polynomial $P(e^{i\theta}, e^{i\phi})$ is said to have *support* S if it can be written in the form

$$(3)\qquad P(e^{i\theta}, e^{i\phi}) = \sum_{k=0}^{n} P_k\, e^{i(s_k\theta + t_k\phi)},$$

i.e., if its coefficients vanish outside the set S. The inner product $(\cdot, \cdot)$ is called *nondegenerate* with respect to S if $(P, P) \neq 0$ for all P of support S except for $P = 0$. We now consider minimization of the functional Φ over the space $L(S)$ consisting of the trigonometric polynomials of support S. The following elementary theorem yields a characterization of minimizing functions in terms of *orthogonality relations* (see [4], e.g.).

THEOREM 1. *Assume the inner product $(\cdot, \cdot)$ to be nondegenerate with respect to the set S. Then the corresponding functional Φ admits exactly one minimizing trigonometric polynomial $P \in L(S)$, characterized by the relation $(P, Q) = Q(0, 0)$ for all Q in $L(S)$.*

Proof. We consider only the case of the right inner product. (The other case is similar.) By definition, $\Phi(P + H) - \Phi(P) = (H, H) + 2 \operatorname{Herm}[(P, H) - H(0, 0)]$. Let us apply this identity to $H = QK$ with $Q \in L(S)$ and $K =$ constant matrix, assuming P to be a minimizing function of support S. Defining $A = (Q, Q)$ and $B = (P, Q) - Q(0, 0)$ one obtains $\tilde{K}AK + 2 \operatorname{Herm}(BK) \geqq 0$, for any K, as a consequence of $\Phi(P + H) \geqq \Phi(P)$. This clearly forces $B = 0$. Conversely, the condition $(P, Q) = Q(0, 0)$ yields $\Phi(Q) - \Phi(P) = (P - Q, P - Q)$, hence $\Phi(P) \leqq \Phi(Q)$, and thus characterizes P as a minimizing function in $L(S)$.

It remains to be shown that the orthogonality equations admit exactly one solution. Define the Hermitian matrix C of order $p(n+1)$ as follows:

$$C = \begin{bmatrix} \Sigma_{0,0} & \Sigma_{s_0-s_1,\,t_0-t_1} & \cdots & \Sigma_{s_0-s_n,\,t_0-t_n} \\ \Sigma_{s_1-s_0,\,t_1-t_0} & \Sigma_{0,0} & \cdots & \Sigma_{s_1-s_n,\,t_1-t_n} \\ \vdots & \vdots & & \vdots \\ \Sigma_{s_n-s_0,\,t_n-t_0} & \Sigma_{s_n-s_1,\,t_n-t_1} & \cdots & \Sigma_{0,0} \end{bmatrix}, \tag{4}$$

with $\Sigma_{s,t} = (4\pi^2)^{-1} \iint e^{-is\theta} e^{-it\phi}\, d\Sigma(\theta, \phi)$ denoting the trigonometric moments of the measure Σ. From (1) and (4) one deduces the identity $(P, Q) = \tilde{X}CY$, with $X = (P_0^T, P_1^T, \cdots, P_n^T)^T$ the coefficient block-vector of P (see (3)) and similarly $Y = (Q_0^T, Q_1^T, \cdots, Q_n^T)^T$. Hence it appears that C is positive definite, owing to non-degeneracy. As a result, the orthogonality equations $(P, Q) = Q(0, 0)$ for the basis of monomials $Q = e^{is\theta} e^{it\phi} I$, which can be written in the form $CX = (I, 0, \cdots, 0)^T$, admit a unique solution. This completes the proof. □

The present paper is exclusively concerned with half-plane supports, i.e., subsets of the *upper half-plane* H consisting of the points $(x, t) \in \mathbb{Z} \times \mathbb{Z}$ such that $t \geqq 0$ for all $s \in \mathbb{Z}$ with $t \geqq 1$ when $s \leqq -1$. It is very interesting to notice that H induces a total ordering of $\mathbb{Z} \times \mathbb{Z}$, compatible with the additive structure, namely the *inverse lexicographic ordering*, given by $(s, t) \leqq (s', t')$ whenever $(s' - s, t' - t) \in H$. (See Helson and Lowdenslager [11].) As for finite subsets $S \subset H$ of particular interest, let us mention the set $S = H_{k,m,j}$ consisting of all points $(s, t) \in H$ satisfying $(s, t) \leqq (m, j)$ and $|s| \leqq k$, where k, m, j are given integers subject to $j \geqq 0$ and $|m| \leqq k$ (cf. [3], [6]). The corresponding matrix C has the half-plane block-Toeplitz structure. Fig. 1 gives the example of such a set S for $k = 4$, $m = 2$, $j = 3$.

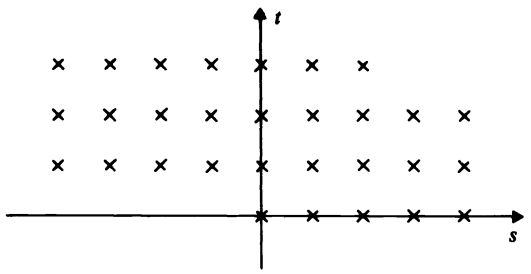


FIG. 1. *The set* $H_{4,2,3}$.

3. One-variable parametric minimization. Before treating the question of two-variable minimization with infinite support, we need some results from the theory of minimization for one-variable polynomials depending on a parameter. The results below are mainly taken from [4]. (See also [21].) Let there be given a Hermitian-valued function $\Delta(\theta, \phi)$ which is nondecreasing with respect to the variable ϕ (almost everywhere in θ). For a nonnegative integer j, let $\Gamma_j(\theta)$ denote the Hermitian block-Toeplitz matrix built on the trigonometric moments of order $0, 1, \cdots, j$ of $\Delta(\theta, \cdot)$, i.e.,

$$\Gamma_j(\theta) = \begin{bmatrix} \Delta_0(\theta) & \Delta_{-1}(\theta) & \cdots & \Delta_{-j}(\theta) \\ \Delta_1(\theta) & \Delta_0(\theta) & \cdots & \Delta_{1-j}(\theta) \\ \vdots & \vdots & & \vdots \\ \Delta_j(\theta) & \Delta_{j-1}(\theta) & \cdots & \Delta_0(\theta) \end{bmatrix}, \tag{5}$$

with $\Delta_s(\theta) = (2\pi)^{-1} \int e^{-is\phi}\, d_\phi \Delta(\theta, \phi)$. For any two matrix polynomials $P(e^{i\theta}, w) = \sum_{t=0}^{j} P_t(e^{i\theta}) w^t$ and $Q(e^{i\theta}, w) = \sum_{t=0}^{j} Q_t(e^{i\theta}) w^t$ of formal degree j in w, with coefficients P_t and Q_t depending on θ, one has the identity

$$\frac{1}{2\pi} \int_0^{2\pi} \tilde{P}(e^{i\theta}, e^{i\phi})\, d_\phi \Delta(\theta, \phi) Q(e^{i\theta}, e^{i\phi}) = \tilde{X}(e^{i\theta}) \Gamma_j(\theta) Y(e^{i\theta}), \tag{6}$$

where X and Y stand for the coefficient block-vectors of P and Q, respectively, i.e., $X = (P_0^T, \cdots, P_j^T)^T$ and $Y = (Q_0^T, \cdots, Q_j^T)^T$.

Assume now $\Gamma_j(\theta)$ is nonsingular (hence positive definite) almost everywhere, and define the matrix polynomial $A_j(e^{i\theta}, w) = \sum_{t=0}^{j} A_{j,t}(e^{i\theta}) w^t$, the coefficients of which are given by

$$(\tilde{A}_{j,0}, \tilde{A}_{j,1}, \cdots, \tilde{A}_{j,j}) = (I, 0, \cdots, 0) \Gamma_j^{-1}. \tag{7}$$

In other words, $A_j(e^{i\theta}, w)$ minimizes the parametric Hermitian-valued functional $\int \tilde{P}(e^{i\theta}, e^{i\phi}) d_\phi \Delta(\theta, \phi) P(e^{i\theta}, e^{i\phi}) - 4\pi \operatorname{Herm} P(e^{i\theta}, 0)$. The minimizing polynomial $B_j(e^{i\theta}, w)$ relative to the dual functional is similarly determined from

$$(B_{j,j}, \cdots, B_{j,1}, B_{j,0}) = (0, \cdots, 0, I) \Gamma_j^{-1}. \tag{8}$$

It is known that both matrices $A_j(e^{i\theta}, 0)$ and $B_j(e^{i\theta}, 0)$ are positive definite, with $A_j(e^{i\theta}, 0) \geqq A_{j-1}(e^{i\theta}, 0)$ and $B_j(e^{i\theta}, 0) \geqq B_{j-1}(e^{i\theta}, 0)$. Note also the identity $B_{j,j}(e^{i\theta}) = A_{j,j}(e^{i\theta})$. Let us now write down the important three-term recurrence relations satisfied by the minimizing polynomials:

$$\begin{aligned} A_{j-1}(e^{i\theta}, w) &= A_j(e^{i\theta}, w) - \hat{B}_j(e^{i\theta}, w) B_{j,0}(e^{i\theta})^{-1} A_{j,j}(e^{i\theta}), \\ B_{j-1}(e^{i\theta}, w) &= B_j(e^{i\theta}, w) - B_{j,j}(e^{i\theta}) A_{j,0}(e^{i\theta})^{-1} \hat{A}_j(e^{i\theta}, w), \end{aligned} \tag{9}$$

with $\hat{A}_j(e^{i\theta}, w) = w^j \tilde{A}_j(e^{i\theta}, 1/\bar{w})$ denoting the reciprocal of $A_j(e^{i\theta}, w)$ and similarly for $\hat{B}_j(e^{i\theta}, w)$. Immediate consequences of (9) are

$$A_{j,0} - A_{j-1,0} = \tilde{B}_{j,j} B_{j,0}^{-1} A_{j,j}, \qquad B_{j,0} - B_{j-1,0} = B_{j,j} A_{j,0}^{-1} \tilde{A}_{j,j}. \tag{10}$$

To progress further we assume the trace of $\Delta_0(\theta)$ and the logarithm of the determinant of $\Gamma_j(\theta)$ to be integrable functions:

$$\operatorname{tr} \Delta_0 \in L_1, \qquad \log \det \Gamma_j \in L_1. \tag{11}$$

Note that (11) implies $\log \det \Gamma_t \in L_1$ for $0 \leqq t \leqq j$. This readily leads to the conclusion that $A_j(e^{i\theta}, 0)^{-1}$ and $B_j(e^{i\theta}, 0)^{-1}$ admit spectral factorizations. Thus one can write

$$A_j(e^{i\theta}, 0)^{-1} = \tilde{M}_j(e^{i\theta}) M_j(e^{i\theta}), \qquad B_j(e^{i\theta}, 0)^{-1} = N_j(e^{i\theta}) \tilde{N}_j(e^{i\theta}), \tag{12}$$

where M_j and N_j are outer matrix-valued functions of class L_2^+ which are uniquely determined within left and right unitary constant factors, respectively (cf. [11], [19]). For reasons that will appear in the sequel we now introduce the *normalized minimizing polynomials* X_j and Y_j, defined as follows:

$$X_j(e^{i\theta}, w) = A_j(e^{i\theta}, w) \tilde{M}_j(e^{i\theta}), \qquad Y_j(e^{i\theta}, w) = \tilde{N}_j(e^{i\theta}) B_j(e^{i\theta}, w). \tag{13}$$

By construction, X_j and Y_j are the reciprocals of the matrix polynomials orthonormal on the unit circle with respect to Δ. They are characterized by the parametric

orthogonality relations

$$\frac{1}{2\pi}\int_0^{2\pi} \tilde{X}_j(e^{i\theta}, e^{i\phi})\, d_\phi\Delta(\theta,\phi)Q(e^{i\theta}, e^{i\phi}) = M_j(e^{i\theta})Q(e^{i\theta}, 0),$$

$$\frac{1}{2\pi}\int_0^{2\pi} Q(e^{i\theta}, e^{i\phi})\, d_\phi\Delta(\theta,\phi)\tilde{Y}_j(e^{i\theta}, e^{i\phi}) = Q(e^{i\theta}, 0)N_j(e^{i\theta}), \tag{14}$$

where $Q(e^{i\theta}, w)$ is any polynomial of formal degree j in w.

Next let us express the recurrence relations (9) in terms of the normalized minimizing polynomials. To that end we introduce the $p\times p$ matrix function $\Omega_j(\theta)$, for $j\geqq 1$, by the equivalent definitions

$$\Omega_j = -\tilde{N}_{j-1}B_{j,j}A_{j,0}^{-1}M_{j-1}^{-1}, \qquad \Omega_j = -N_{j-1}^{-1}B_{j,0}^{-1}A_{j,j}\tilde{M}_{j-1}, \tag{15}$$

where the argument θ is omitted for convenience. (The consistency of (15) is a straightforward consequence of (10) and (12) together with $A_{j,j} = B_{j,j}$.) From (10), (12) and (15) one deduces

$$I-\tilde{\Omega}_j\Omega_j = \tilde{M}_{j-1}^{-1}\tilde{M}_jM_jM_{j-1}^{-1}, \qquad I-\Omega_j\tilde{\Omega}_j = N_{j-1}^{-1}N_j\tilde{N}_j\tilde{N}_{j-1}^{-1}, \tag{16}$$

which shows that $\Omega_j(\theta)$ is strictly contractive almost everywhere. The normalized version of (9) is easily found to be

$$X_{j-1} = X_j\tilde{M}_j^{-1}\tilde{M}_{j-1} + \hat{Y}_jN_j^{-1}N_{j-1}\Omega_j, \qquad Y_{j-1} = \tilde{N}_{j-1}\tilde{N}_j^{-1}Y_j + \Omega_jM_{j-1}M_j^{-1}\hat{X}_j, \tag{17}$$

with $\hat{X}_j$ and $\hat{Y}_j$ the reciprocals of X_j and Y_j. Applying (14), one obtains from (17) useful expressions for $\Omega_j(\theta)$, namely

$$\Omega_j(\theta) = \left[\frac{1}{2\pi}\int e^{-ij\phi}Y_{j-1}(e^{i\theta}, e^{i\phi})\, d_\phi\Delta(\theta,\phi)\right]M_{j-1}(e^{i\theta})^{-1},$$

$$\Omega_j(\theta) = N_{j-1}(e^{i\theta})^{-1}\left[\frac{1}{2\pi}\int e^{-ij\phi}\, d_\phi\Delta(\theta,\phi)X_{j-1}(e^{i\theta}, e^{i\phi})\right]. \tag{18}$$

In order to write (17) in a very convenient compact form, let us introduce the $2p\times 2p$ matrix function $U_j(e^{i\theta})$ as

$$U_j = \begin{bmatrix} I & -\Omega_j \\ -\tilde{\Omega}_j & I \end{bmatrix}\begin{bmatrix} \tilde{N}_{j-1}\tilde{N}_j^{-1} & 0 \\ 0 & M_{j-1}M_j^{-1} \end{bmatrix}. \tag{19}$$

In view of (16) it appears that U_j is *J-unitary*, in the sense that it satisfies $\tilde{U}_jJU_j = U_jJ\tilde{U}_j = J$, where J denotes the diagonal matrix $I \dot{+} (-I)$. It is then easily verified that the inverse form of (17) can be written as

$$[\hat{Y}_j \quad X_j] = [w\hat{Y}_{j-1} \quad X_{j-1}]U_j, \tag{20}$$

owing to the property $U_j^{-1} = J\tilde{U}_jJ$. Using (20) in an inductive manner one immediately obtains, by the J-unitarity of U_j, the identity

$$X_j(e^{i\theta}, w)\hat{X}_j(e^{i\theta}, w) = \hat{Y}_j(e^{i\theta}, w)Y_j(e^{i\theta}, w). \tag{21}$$

Let us finally recall how the moments $\Delta_0, \Delta_1, \cdots, \Delta_j$ can be reconstructed from X_j and Y_j. The minimizing polynomials are known to be nonsingular in the closed unit disk $|w|\leqq 1$. Furthermore, the inverse of X_j satisfies

$$\frac{1}{2\pi}\int e^{-it\phi}\tilde{X}_j^{-1}X_j^{-1}\, d\phi = \Delta_t - \delta_{t,j+1}N_j\Omega_{j+1}M_j, \tag{22}$$

for $t = 0, 1, \cdots, j+1$, where δ is the Kronecker symbol. In view of (21) a similar result holds for Y_j. (In fact, (22) is explicitly given in [4] only for $t \leqq j$, but the case $t = j+1$ also follows without difficulty from the results of [4].)

4. Minimization for infinite supports of finite width. From now on (with an exception in the beginning of § 7) we assume the p-dimensional measure $\Sigma(\theta, \phi)$ to be absolutely continuous with respect to θ. Thus one can write

$$\Sigma(\theta, \phi) = \int_0^\theta \Delta(\alpha, \phi)\, d\alpha, \tag{23}$$

where the Hermitian-valued function $\Delta(\alpha, \phi)$ is nondecreasing in ϕ. Henceforth the integrability assumptions (11) on Δ are replaced by the corresponding more restrictive assumptions of boundedness, namely

$$\operatorname{tr} \Delta_0 \in L_\infty, \qquad \det \Gamma_j^{-1} \in L_\infty, \tag{24}$$

with Γ_j as in (5), for a given integer $j \geqq 0$. (This implies $\det \Gamma_t^{-1} \in L_\infty$ for $t < j$.) As a consequence of the boundedness of Γ_j, the right member of (6) belongs to the Lebesgue space L_1, provided the coefficient functions $P_t(e^{i\theta})$ and $Q_t(e^{i\theta})$ are in L_2. In this case the inner product $(P, Q)_r$ is well defined and can be written as the iterated integral

$$(P, Q)_r = \frac{1}{4\pi^2} \int d\theta \int \tilde{P}(e^{i\theta}, e^{i\phi})\, d_\phi \Delta(\theta, \phi) Q(e^{i\theta}, e^{i\phi}). \tag{25}$$

Since P and Q are assumed to have formal degree j in w, application of (22) to the ϕ-integral in (25) yields

$$(P, Q)_r = \frac{1}{4\pi^2} \int d\theta \int \tilde{P} \tilde{X}_j^{-1} X_j^{-1} Q\, d\phi. \tag{26}$$

Expressions similar to (25) and (26) hold for the left inner product $(P, Q)_l$. Let us mention a useful Schwarz inequality on both inner products, namely

$$\|(P, Q)\| \leqq p(j+1) \|\operatorname{tr} \Delta_0\|_\infty \|P\|_2 \|Q\|_2. \tag{27}$$

Here and in the sequel $\|\cdot\|$ denotes the spectral norm, $\|\cdot\|_\infty$ is the usual sup norm and $\|\cdot\|_2$ is the matrix L_2-norm (i.e., $4\pi^2 \|P\|_2^2 = \iint \operatorname{tr} (\tilde{P}P)\, d\theta\, d\phi$). The proof of (27) from (25) is elementary and left to the reader.

Let S be any subset of the integer plane $\mathbb{Z} \times \mathbb{Z}$. A two-variable matrix function $P(e^{i\theta}, e^{i\phi})$ is said to be of *class* $L_q(S)$ if its elements belong to the Lebesgue space L_q, for a given $q \geqq 1$, and if they admit the set S as Fourier support. In the present section we consider the problem of minimizing the functionals Φ_r and Φ_l in the class $L_2(S)$, for a given subset S of the *horizontal strip*

$$H_j = \{(s, t) \in \mathbb{Z} \times \mathbb{Z} : (0, 0) \leqq (s, t) < (\infty, j)\}, \tag{28}$$

where the symbols $\leqq$ and $<$ are used for the inverse lexicographic ordering (see the end of § 2). Thus H_j contains all points (s, t) of the upper half-plane H subject to $t \leqq j$. Fig. 2 shows the case $j = 3$.

As pointed out above, the functionals Φ_r and Φ_l are well defined over $L_2(S)$ with $S \subset H_j$. Before characterizing the minimizing functions (Theorem 4) let us emphasize the main properties of the polynomials X_j and Y_j that result from (24).

THEOREM 2. *The normalized minimizing polynomial $X_j(e^{i\theta}, w)$ belongs to $L_\infty(H_j)$ and its inverse $X_j(e^{i\theta}, w)^{-1}$ belongs to $L_2(H)$. Furthermore, $X_j(e^{i\theta}, w)^{-1}$ is essentially*

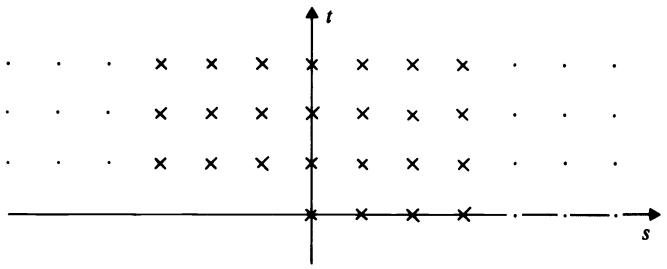


FIG. 2. *The strip* H_3.

bounded in θ, *uniformly on every closed disk* $|w| \leqq r < 1$. *The same properties hold for* $Y_j(e^{i\theta}, w)$.

Proof. By assumption, Γ_j and Γ_j^{-1} have bounded entries, so that the functions $A_{j,t}$ defined from (7) belong to L_∞. On the other hand, $A_{j,0} \geqq A_{0,0} = \Delta_0^{-1}$ implies $A_{j,0}^{-1} \in L_\infty$. Hence one has $M_j \in L_\infty^+$ and $M_j^{-1} \in L_\infty^+$ in (12). This clearly leads to the first conclusion $X_j \in L_\infty(H_j)$, in view of $X_{j,0} = M_j^{-1}$. The second assertion, $X_j^{-1} \in L_2(H)$, directly follows from $X_{j,0}^{-1} \in L_\infty^+$ together with the fact that X_j is nonsingular in $|w| \leqq 1$ and satisfies $\int \tilde{X}_j^{-1} X_j^{-1}\, d\phi = 2\pi\, \Delta_0 \in L_\infty$ (see (22)). Applying Poisson's inequality to the last identity one obtains

$$\tilde{X}_j(e^{i\theta}, w)^{-1} X_j(e^{i\theta}, w)^{-1} \leqq \frac{1+r}{1-r} \Delta_0(\theta) \tag{29}$$

for $|w| \leqq r < 1$, which proves the third statement. The properties of Y_j are established in the same manner. □

From the normalized minimizing polynomials X_j, Y_j and the spectral factors M_j, N_j, let us define both polynomials

$$F_j(e^{i\theta}, w) = X_j(e^{i\theta}, w) \tilde{M}_j(0)^{-1}, \qquad G_j(e^{i\theta}, w) = \tilde{N}_j(0)^{-1} Y_j(e^{i\theta}, w). \tag{30}$$

THEOREM 3. *The polynomials* F_j *and* G_j *minimize the functionals* Φ_r *and* Φ_l, *respectively, over the space* $L_2(H_j)$.

Proof. Theorem 2 shows that F_j and G_j belong to $L_\infty(H_j)$. On the other hand, θ-integration of the first equation (14) gives $(F_j, Q)_r = Q(0, 0)$, for all $Q \in L_2(H_j)$, in view of (25). By the same argument as in the proof of Theorem 1, this yields the desired property $\Phi_r(F_j) \leqq \Phi_r(Q)$. The dual result $\Phi_l(G_j) \leqq \Phi_l(Q)$ whenever $Q \in L_2(H_j)$ is proved analogously. □

The general result of Theorem 4 below about minimization in any class $L_2(S)$ follows from considering two particularly simple situations, namely the finite case (Theorem 1) and the extremal case (Theorem 3). In fact, the first result provides approximants for the general problem while the second result is used to prove convergence of these approximants.

THEOREM 4. *Let* S *be any subset of the horizontal strip* H_j *containing the origin* $(0, 0)$. *Then the functional* Φ *admits a unique minimizing function* $P(e^{i\theta}, w) \in L_2(S)$, *characterized by the orthogonality relations* $(P, Q) = Q(0, 0)$ *for all* $Q \in L_2(S)$.

Proof. To be specific we consider the case of the right inner product. Note first that the situation where S is finite is covered by Theorem 1. (The nondegeneracy property is obvious from (26).) Next we shall establish the existence of P for an infinite support S. Let $S_0 \subset S_1 \subset \cdots \subset S_n \subset \cdots \subset S$ be an ascending chain of finite subsets S_n of S, with $(0, 0) \in S_0$ and $\bigcup_{n=0}^{\infty} S_n = S$. Let P_n denote the minimizing trigonometric polynomial in

$L(S_n)$. The argument consists in showing the existence of the function $P \in L_2(S)$ given by

$$P(\cdot, w) = \operatorname*{l.i.m.}_{n\to\infty} P_n(\cdot, w). \tag{31}$$

Theorem 1 yields $(P_m - P_n, P_m - P_n) = P_m(0,0) - P_n(0,0)$ for $m \geqq n$; hence, by use of (26) and Poisson's inequality,

$$\|X_j(\cdot, w)^{-1}[P_m(\cdot, w) - P_n(\cdot, w)]\|_2^2 \leqq \frac{1+r}{1-r} \operatorname{tr}[P_m(0,0) - P_n(0,0)], \tag{32}$$

in the disk $|w| \leqq r < 1$. By Theorem 2, the coefficients $X_{j,t}$ are bounded in θ, so that (32) readily leads to

$$\|P_m(\cdot, w) - P_n(\cdot, w)\|_2^2 \leqq c(r) \operatorname{tr}[P_m(0,0) - P_n(0,0)], \tag{33}$$

for a suitable constant $c(r)$. Note that (33) remains valid in any disk $|w| \leqq r < \infty$, within adjustment of $c(r)$, because $P_m - P_n$ is a polynomial of fixed degree j in w.

From Theorem 3 one immediately deduces $\Phi(F_j) \leqq \Phi(P_n)$, i.e., $P_n(0,0) \leqq F_j(0,0)$, for all n. As a result, the sequence of numbers $\operatorname{tr} P_n(0,0)$ is bounded, hence convergent since it is nondecreasing. Then, in view of (33), the Cauchy criterion shows that the limit function (31) exists and belongs to $L_2(S)$. It is easily seen that P satisfies the orthogonality relations in $L_2(S)$. Indeed, Theorem 1 yields $(P_n, Q_k) = Q_k(0,0)$ for all $Q_k \in L(S_k)$ with $k \leqq n$. Let $n \to \infty$ for a given Q_k. Owing to (31) and (27) one obtains $(P, Q_k) = Q_k(0,0)$. Hence the desired property follows from the fact that any function $Q \in L_2(S)$ is the limit in the mean of an appropriate sequence of trigonometric polynomials $Q_k \in L(S_k)$. Next the equivalence between minimality and orthogonality is proved by the same argument as in Theorem 1. The uniqueness of the minimizing function is then immediate. □

THEOREM 5. *Let there be given an ascending chain $S_0 \subset S_1 \subset S_2 \subset \cdots \subset H_j$ or a descending chain $H_j \supset S_0 \supset S_1 \supset S_2 \supset \cdots$. Then, for $n \to \infty$, the minimizing function of support S_n converges in the mean to the minimizing function of support $\bigcup_{n=0}^{\infty} S_n$ in the first case and $\bigcap_{n=0}^{\infty} S_n$ in the second case.*

Proof. The argument is exactly the same as in the proof of Theorem 4 for the ascending situation. The second case is quite similar. Details are omitted. □

Let us conclude with a remark. By Theorem 4 the minimizing function $P(e^{i\theta}, w)$ in any class $L_2(S)$ satisfies $P(0,0) = (P, P) = -\Phi(P)$. This shows that $P(0,0)$ is a positive definite Hermitian matrix which is monotonically increasing with respect to the support S (in the sense that $P_1(0,0) \leqq P_2(0,0)$ when $S_1 \subset S_2$).

5. Minimization for lexicographic interval supports. In this section we analyze in great detail the minimizing functions of support $H_{m,j}$, where the set $H_{m,j}$ is an *interval* with respect to the inverse lexicographic order (see § 2), i.e.,

$$H_{m,j} = \{(s,t) \in \mathbb{Z} \times \mathbb{Z} : (0,0) \leqq (s,t) \leqq (m,j)\}, \tag{34}$$

for a fixed nonnegative integer j and any integer m. In other words, the point (s,t) belongs to $H_{m,j}$ if and only if both (s,t) and $(m-s, j-t)$ belong to the upper half-plane H. Fig. 3 shows the example of (34) for $m = 2$, $j = 3$. Of course $H_{m,j}$ is infinite when $j \geqq 1$, while $H_{m,0}$ is the interval $[0, m]$ of the s-axis (which is empty in case $m < 0$).

Let $F_{m,j}(e^{i\theta}, w)$ and $G_{m,j}(e^{i\theta}, w)$ denote the minimizing polynomials in the space $L_2(H_{m,j})$ relative to the functionals Φ_r and Φ_l, respectively (see Theorem 4). Application

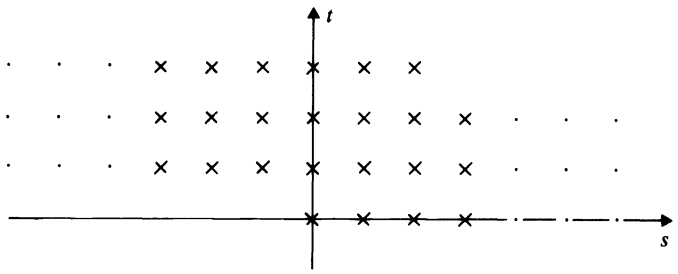


FIG. 3. *The interval* $H_{2,3}$.

of Theorems 3 and 5 directly yields

$$\text{(35)} \qquad \underset{m\to+\infty}{\text{l.i.m.}} F_{m,j} = F_j, \qquad \underset{m\to-\infty}{\text{l.i.m.}} F_{m,j} = F_{j-1},$$

and similarly $G_{m,j} \to G_j$ when $m \to +\infty$ and $G_{m,j} \to G_{j-1}$ when $m \to -\infty$, where F_t and G_t are defined as in (30). On the other hand, let $F_{k,m,j}$ and $G_{k,m,j}$ be the trigonometric polynomials that minimize Φ_r and Φ_l, respectively, for the finite support $H_{k,m,j}$ indicated in § 2. Then it follows from Theorem 5 that $F_{k,m,j}$ converges in the mean to $F_{m,j}$ when k tends to infinity. Similarly $G_{k,m,j} \to G_{m,j}$ when $k \to \infty$.

It is often more convenient to consider the *normalized minimizing functions* $X_{m,j}$ and $Y_{m,j}$ given by

$$\text{(36)} \qquad X_{m,j}(e^{i\theta}, w) = F_{m,j}(e^{i\theta}, w)M_{m,j}, \qquad Y_{m,j}(e^{i\theta}, w) = N_{m,j}G_{m,j}(e^{i\theta}, w),$$

where the constant $p \times p$ matrices $M_{m,j}$ and $N_{m,j}$ are chosen so as to satisfy $M_{m,j}\tilde{M}_{m,j} = F_{m,j}(0,0)^{-1}$ and $\tilde{N}_{m,j}N_{m,j} = G_{m,j}(0,0)^{-1}$. (This makes sense because $F_{m,j}(0,0)$ and $G_{m,j}(0,0)$ are positive definite; see the remark at the end of § 4.) When expressed in terms of the normalized minimizing functions, the orthogonality relations of Theorem 4 are

$$\text{(37)} \qquad (X_{m,j}, Q)_r = X_{m,j}(0,0)^{-1}Q(0,0), \qquad (Y_{m,j}, Q)_l = Q(0,0)\,Y_{m,j}(0,0)^{-1},$$

for Q varying over $L_2(H_{m,j})$. Note that $X_{m,j}$ and $Y_{m,j}$ are only defined within arbitrary right and left unitary factors, respectively, but they uniquely determine the minimizing polynomials as $F_{m,j} = X_{m,j}\tilde{X}_{m,j}(0,0)$ and $G_{m,j} = \tilde{Y}_{m,j}(0,0)\,Y_{m,j}$. Let us incidentally point out that the class $L_2(H_{m,j})$ functions $X^*_{m,j}(e^{i\theta}, e^{i\phi}) = e^{im\theta}\,e^{ij\phi}\tilde{X}_{m,j}(e^{i\theta}, e^{i\phi})$ are pairwise orthonormal with respect to the inner product $(\cdot,\cdot)_l$ in the sense that they satisfy $(X^*_{m,j}, X^*_{n,k})_l = \delta_{m,n}\,\delta_{j,k}I$ for all m, j, n, k. Similarly the functions $Y^*_{m,j} = e^{im\theta}\,e^{ij\phi}\tilde{Y}_{m,j}$ are orthonormal with respect to the inner product $(\cdot,\cdot)_r$.

5.1. Construction via generalized Schur representation. For $m \geqq 0$ the functions $X_{m,0}$ and $Y_{m,0}$ simply are the reciprocals of the left and right orthonormal polynomials of degree m associated with the weight function $\Delta_0(\theta)$ (see [4], [21]). Henceforth we treat the more difficult problem $j \geqq 1$. The expression given in Theorem 7 below for the normalized minimizing polynomials $X_{m,j}$ and $Y_{m,j}$ is crucially based upon the following result concerning the generalized Schur representation [7] of the contractive matrix function $e^{-im\theta}\Omega_j(\theta)$, with Ω_j defined as in (13).

LEMMA 6. *There exist eight $p\times p$ matrix functions $A_{m,j}(e^{i\theta}),\cdots,D_{m,j}(e^{i\theta})$, $P_{m,j}(e^{i\theta}),\cdots,S_{m,j}(e^{i\theta})$ of class L_2^+ satisfying*

$$\begin{bmatrix} I & -e^{-im\theta}\Omega_j \\ -e^{im\theta}\tilde{\Omega}_j & I \end{bmatrix}\begin{bmatrix} \tilde{A}_{m,j} & \tilde{B}_{m,j} \\ C_{m,j} & D_{m,j} \end{bmatrix}=\begin{bmatrix} P_{m,j} & -e^{i\theta}Q_{m,j} \\ -e^{-i\theta}\tilde{R}_{m,j} & \tilde{S}_{m,j} \end{bmatrix}, \tag{38}$$

together with $A_{m,j}(0)P_{m,j}(0)=S_{m,j}(0)D_{m,j}(0)=I$. These functions are uniquely determined except for normalization corresponding to postmultiplication of (38) *by the direct sum of any two constant unitary matrices of order p. The functions $A_{m,j}^{-1}$ and $D_{m,j}^{-1}$ necessarily belong to the class L_∞^+. On the other hand, the second factor in the left member of* (38) *is J-unitary.*

Proof. This is part of [7, Thm. 1] (which has intimate connections with some results by Adamjan, Arov and Krein [1]). □

Let us introduce the J-unitary matrix function $U_{m,j}(e^{i\theta})$ obtained by a simple transformation of the second factor in (38), that is,

$$U_{m,j}(e^{i\theta})=\begin{bmatrix} \tilde{A}_{m,j}(e^{i\theta}) & -e^{im\theta}\tilde{B}_{m,j}(e^{i\theta}) \\ -e^{-im\theta}C_{m,j}(e^{i\theta}) & D_{m,j}(e^{i\theta}) \end{bmatrix}. \tag{39}$$

THEOREM 7. *The normalized minimizing polynomials $X_{m,j}(e^{i\theta},w)$ and $Y_{m,j}(e^{i\theta},w)$ of support $H_{m,j}$ are given in terms of the class L_2^+ functions $A_{m,j}(e^{i\theta})$, $B_{m,j}(e^{i\theta})$, $C_{m,j}(e^{i\theta})$, $D_{m,j}(e^{i\theta})$ and of the polynomials $X_{j-1}(e^{i\theta},w)$, $Y_{j-1}(e^{i\theta},w)$ defined in* § 3 *by the expression*

$$[\hat{Y}_{m,j}\quad X_{m,j}]=[w\hat{Y}_{j-1}\quad X_{j-1}]U_{m,j}, \tag{40}$$

with $\hat{Y}_{m,j}(e^{i\theta},w)=w^j\tilde{Y}_{m,j}(e^{i\theta},1/\bar{w})$ denoting the reciprocal of $Y_{m,j}(e^{i\theta},w)$ and similarly $\hat{Y}_{j-1}(e^{i\theta},w)=w^{j-1}\tilde{Y}_{j-1}(e^{i\theta},1/\bar{w})$.

Proof. Note first that, in view of Theorem 2, the functions $X_{m,j}$ and $Y_{m,j}$ determined from (40) actually belong to $L_2(H_{m,j})$. To prove the statement one has to check (37) for the monomials $Q(e^{i\theta},w)=e^{is\theta}w^tI$ with $(s,t)\in H_{m,j}$. It clearly appears from (17) that the required conditions are satisfied for $1\leqq t\leqq j-1$. Next one shows, by use of (17) and (18), that the conditions for $t=0$ and $t=j$ are exactly those of Lemma 6. The details are left to the reader (cf. [6] for the scalar case). □

For future use it is convenient to give a meaning to (40) in the case $j=0$, $m\geqq 0$. In fact, there exists a well-defined J-unitary matrix $U_{m,0}$ (having properties similar to those of $U_{m,j}$ for $j\geqq 1$) such that (40) holds true with the interpretation $w\hat{Y}_{-1}=I$, $X_{-1}=0$ (see [5]). It is also interesting to express $X_{m,j}$ and $Y_{m,j}$ in terms of X_j and Y_j, instead of X_{j-1} and Y_{j-1}. To that end let us define the $2p\times 2p$ matrix function $V_{m,j}$, closely related to the dual of $U_{m,j}$ in the sense of [7], by the following expression:

$$V_{m,j}=\begin{bmatrix} N_j^{-1}N_{j-1} & 0 \\ 0 & \tilde{M}_j^{-1}\tilde{M}_{j-1} \end{bmatrix}\begin{bmatrix} P_{m,j} & e^{i(m+1)\theta}Q_{m,j} \\ e^{-i(m+1)\theta}\tilde{R}_{m,j} & \tilde{S}_{m,j} \end{bmatrix}, \tag{41}$$

where P, Q, R, S are as in Lemma 6 and M, N as in § 3. In view of (19) one has $V_{m,j}=U_j^{-1}U_{m,j}$, which shows that $V_{m,j}$ is J-unitary. From (20) and (40) one then deduces

$$[\hat{Y}_{m,j}\quad X_{m,j}]=[\hat{Y}_j\quad X_j]V_{m,j}. \tag{42}$$

As proved in [7], for a suitable normalization the matrix function $U_{m,j}$ converges in the mean to I when $m\to-\infty$ and to U_j when $m\to+\infty$. (The meaning of $U_{m,0}\to U_0$ for $m\to\infty$ can be found in [5].) Equivalently, $V_{m,j}\to U_j^{-1}$ when $m\to-\infty$ and $V_{m,j}\to I$ when $m\to+\infty$. These results provide an alternative proof of (35).

Let us finally mention some useful identities. Using the J-unitarity of $U_{m,j}$ or of $V_{m,j}$ one obtains from (40) or (42), together with (21),

$$X_{m,j}(e^{i\theta}, w)\hat{X}_{m,j}(e^{i\theta}, w) = \hat{Y}_{m,j}(e^{i\theta}, w)Y_{m,j}(e^{i\theta}, w). \tag{43}$$

Let $X_{m,j}^{(t)}(e^{i\theta})$ and $Y_{m,j}^{(t)}(e^{i\theta})$ denote the coefficients of w^t in the polynomials $X_{m,j}(e^{i\theta}, w)$ and $Y_{m,j}(e^{i\theta}, w)$, respectively. Then (40) immediately yields

$$\begin{bmatrix} \tilde{Y}_{m,j}^{(0)} & X_{m,j}^{(j)} \\ \tilde{Y}_{m,j}^{(j)} & X_{m,j}^{(0)} \end{bmatrix} = \begin{bmatrix} \tilde{N}_{j-1}^{-1} & 0 \\ 0 & M_{j-1}^{-1} \end{bmatrix} U_{m,j}. \tag{44}$$

5.2. Stability and related results. In this section the minimizing polynomials of interval support are shown to enjoy stability properties similar to those mentioned in Theorem 2. In addition, partial reconstruction of the trigonometric moments of Σ is obtained. (See [16] for the scalar case.)

THEOREM 8. *The inverse $X_{m,j}^{-1}$ of the normalized minimizing polynomial $X_{m,j}$ belongs to the space $L_2(H)$, where H is the upper half-plane. Furthermore, $X_{m,j}(e^{i\theta}, w)^{-1}$ is essentially bounded in θ, uniformly on every closed disk $|w| \leqq r < 1$. The same properties hold for $Y_{m,j}$.*

Proof. Let us write the second block equation of (40) in the form $X_{j-1}^{-1}X_{m,j}D_{m,j}^{-1} = I - \Psi$, where Ψ is defined as follows:

$$\Psi(e^{i\theta}, w) = e^{im\theta}wX_{j-1}^{-1}\hat{Y}_{j-1}\tilde{B}_{m,j}D_{m,j}^{-1}. \tag{45}$$

In view of (21) and $\tilde{D}_{m,j}D_{m,j} - B_{m,j}\tilde{B}_{m,j} = I$ (resulting from the J-unitarity of $U_{m,j}$), one has $I - \tilde{\Psi}\Psi = \tilde{D}_{m,j}^{-1}D_{m,j}^{-1}$ for $w = e^{i\phi}$. As a consequence, $\Psi(e^{i\theta}, w)$ is a matrix-valued Schur function of w, in the strict sense, so that $I - \Psi$ is nonsingular for $|w| \leqq 1$ (almost everywhere in θ). Hence $X_{m,j}(e^{i\theta}, w)^{-1}$ is analytic in the closed unit disk $|w| \leqq 1$. On the other hand, straightforward computation yields

$$\tilde{X}_{m,j}^{-1}X_{m,j}^{-1} = \tilde{X}_{j-1}^{-1}[(I-\Psi)^{-1} + (I-\tilde{\Psi})^{-1} - I]X_{j-1}^{-1}, \tag{46}$$

on the unit circle $w = e^{i\phi}$. Now it is easily seen, by use of (45) and (21), that the ϕ-integrals of all functions $\tilde{X}_{j-1}^{-1}\Psi^k X_{j-1}^{-1}$ vanish for $k \geqq 1$. Thus (46) and (22) imply

$$\int_0^{2\pi} \tilde{X}_{m,j}^{-1}X_{m,j}^{-1}\,d\phi = \int_0^{2\pi} \tilde{X}_{j-1}^{-1}X_{j-1}^{-1}\,d\phi = 2\pi\Delta_0(\theta). \tag{47}$$

As a result, the entries of $X_{m,j}(e^{i\theta}, e^{i\phi})^{-1}$ are square-integrable. From the fact that $X_{m,j}^{-1}$ is analytic in $|w| \leqq 1$ and from $X_{m,j}(e^{i\theta}, 0)^{-1} = D_{m,j}(e^{i\theta})^{-1}M_{j-1}(e^{i\theta}) \in L_\infty^+$, this implies that $X_{m,j}^{-1}$ belongs to $L_2(H)$. Note finally that application of the Poisson inequality to (47) yields (29), with $X_{m,j}$ substituted for X_j. This concludes the proof of the assertions concerning $X_{m,j}$. The properties of $Y_{m,j}$ are established by similar arguments. □

THEOREM 9. *For all points (s, t) in the interval $H_{m,j}$ the trigonometric moments $\Sigma_{s,t}$ relative to the measure Σ coincide with the corresponding moments relative to the Hermitian-valued weight function*

$$W_{m,j}(\theta, \phi) = \tilde{X}_{m,j}(e^{i\theta}, e^{i\phi})^{-1}X_{m,j}(e^{i\theta}, e^{i\phi})^{-1}, \tag{48}$$

which is equivalently given by $W_{m,j} = Y_{m,j}^{-1}\tilde{Y}_{m,j}^{-1}$ (cf. (43)). In other words, the inner products (1) relative to both measures $d\Sigma$ and $W_{m,j}\,d\theta\,d\phi$ coincide over the space $L_2(H_{m,j})$.

Proof. This result appears as an immediate consequence of the following identity, which is valid for all $t \in [0, j]$:

$$\frac{1}{2\pi}\int_0^{2\pi} e^{-it\phi} W_{m,j}\, d\phi = \Delta_t - \delta_{t,j}\, e^{i(m+1)\theta} N_{j-1} Q_{m,j} D_{m,j}^{-1} M_{j-1}, \tag{49}$$

where $D_{m,j}$ and $Q_{m,j}$ are as in Lemma 6. It is indeed clear that multiplying (49) by $e^{-is\theta}$ and integrating in θ produces equality between the (s, t)-moments relative to $W_{m,j}\, d\theta\, d\phi$ and to $d\Sigma$ whenever $(0, 0) \leqq (s, t) \leqq (m, j)$.

Let us now derive (49) by use of (46). Observe that the ϕ-integrals of the functions $e^{-it\phi}\tilde{X}_{j-1}^{-1}\tilde{\Psi}^k X_{j-1}^{-1}$ and $e^{-it\phi}\tilde{X}_{j-1}^{-1}\Psi^k X_{j-1}^{-1}$ vanish for $k \geqq 1$ and for $k \geqq 2$, respectively. Hence the left member of (49) is found to be equal to

$$\frac{1}{2\pi}\int e^{-it\phi}\tilde{X}_{j-1}^{-1} X_{j-1}^{-1}\, d\phi + \delta_{t,j}\, e^{im\theta} N_{j-1}\tilde{B}_{m,j} D_{m,j}^{-1} M_{j-1}. \tag{50}$$

Now Lemma 6 gives $\tilde{B}_{m,j} = e^{-im\theta}\Omega_j D_{m,j} - e^{i\theta} Q_{m,j}$. Thus, in view of (22), the right member of (49) coincides with (50), which concludes the proof. □

From (44) one deduces the following matrix inequalities, which are related to stability properties (see Theorem 11 below):

$$X_{m,j}^{(j)}\tilde{X}_{m,j}^{(j)} \leqq \tilde{Y}_{m,j}^{(0)} Y_{m,j}^{(0)}, \qquad \tilde{Y}_{m,j}^{(j)} Y_{m,j}^{(j)} \leqq X_{m,j}^{(0)}\tilde{X}_{m,j}^{(0)}. \tag{51}$$

5.3. Three-term recurrence relations. From the matrix-valued functions $A_{m,j}$, $B_{m,j}$, $C_{m,j}$, $D_{m,j}$ occurring in (39), define the $p \times p$ constant matrix $E_{m,j}$ to be

$$E_{m,j} = B_{m,j}(0) A_{m,j}(0)^{-1} = D_{m,j}(0)^{-1} C_{m,j}(0). \tag{52}$$

In the sequel $E_{m,j}$ is referred to as a *Schur parameter.* (See [7] for $j \geqq 1$ and [5] for $j = 0$. By convention, $E_{m,0} = 0$ when $m \leqq 0$.) It turns out that $E_{m,j}$ is *strictly contractive*, i.e., $\|E_{m,j}\| < 1$. Incidentally, we mention the square-summability property $\sum_{m=-\infty}^{+\infty} \|E_{m,j}\|^2 < \infty$ (cf. [5], [7]). From $E_{m,j}$ let us construct the $2p \times 2p$ matrix function

$$Z_{m,j}(e^{i\theta}) = \begin{bmatrix} (I - \tilde{E}_{m,j}E_{m,j})^{-1/2} & -e^{im\theta}\tilde{E}_{m,j}(I - E_{m,j}\tilde{E}_{m,j})^{-1/2} \\ -e^{-im\theta}E_{m,j}(I - \tilde{E}_{m,j}E_{m,j})^{-1/2} & (I - E_{m,j}\tilde{E}_{m,j})^{-1/2} \end{bmatrix}. \tag{53}$$

It is readily verified that $Z_{m,j}$ is Hermitian and J-unitary. As shown in [7], the matrices (39) obey the recurrence relation

$$U_{m,j}(e^{i\theta}) = U_{m-1,j}(e^{i\theta}) Z_{m,j}(e^{i\theta}). \tag{54}$$

As an immediate consequence of (40) and (54) one has the following important result, which is very interesting from the application viewpoint (cf. Marzetta [16] and the authors [6]).

THEOREM 10. *The normalized minimizing polynomials $X_{m,j}$ and $Y_{m,j}$ are deduced from $X_{m-1,j}$ and $Y_{m-1,j}$ by means of the three-term recurrence relation built on the Schur parameter $E_{m,j}$; i.e.,*

$$[\hat{Y}_{m,j} \quad X_{m,j}] = [\hat{Y}_{m-1,j} \quad X_{m-1,j}] Z_{m,j}. \tag{55}$$

Remarks. Let us first briefly indicate how (55) can be directly obtained in terms of the minimizing polynomials $F_{m,j}(e^{i\theta}, w)$ and $G_{m,j}(e^{i\theta}, w)$. In fact, from the orthogonality relations of Theorem 4 one readily deduces

$$F_{m-1,j} = F_{m,j} + e^{im\theta}\hat{G}_{m,j} K_{m,j}, \qquad G_{m-1,j} = G_{m,j} + e^{im\theta} L_{m,j}\hat{F}_{m,j}, \tag{56}$$

for well-defined constant matrices $K_{m,j}$ and $L_{m,j}$. Applying then (56) in an inductive manner, one obtains an identity of the type $F_{n,j} = F_{n+k,j}\tilde{A}_{n,j,k} + e^{i(n+1)\theta}\hat{G}_{n+k,j}B_{n,jk}$, where $A_{n,j,k}$ and $B_{n,j,k}$ are matrix polynomials of formal degree $k-1$ in $e^{i\theta}$. A similar identity holds for $G_{n,j}$, involving polynomials $C_{n,j,k}$ and $D_{n,j,k}$. For $k \to \infty$ this leads to an equation of the form (42), since $F_{n+k,j} \to F_j$ and $G_{n+k,j} \to G_j$, provided the sequences $A_{n,j,k}, \cdots, D_{n,j,k}$ are convergent in the space L_2^+. This argument provides a heuristic explanation of the expressions (40) and (42) for the minimizing functions of interval support.

From a computational viewpoint it is interesting to note that the matrices $K_{m,j}$ and $L_{m,j}$ occurring in (56) are determined in terms of the minimizing functions of support $H_{m-1,j}$ via the formulas

$$K_{m,j} = (e^{im\theta}w^j I, F_{m-1,j})_r, \qquad L_{m,j} = (e^{im\theta}w^j I, G_{m-1,j})_l. \tag{57}$$

As for the Schur parameters $E_{m,j}$, they are given both by $E_{m,j} = \tilde{M}_{m,j}\tilde{K}_{m,j}N_{m,j}^{-1}$ and by $E_{m,j} = M_{m,j}^{-1}\tilde{L}_{m,j}\tilde{N}_{m,j}$, where $M_{m,j}$ and $N_{m,j}$ are as in (36).

Let us now emphasize an important fact about the recurrence relations of Theorem 10: the stability properties of the minimizing functions with interval support are hereditary (provided only the Schur parameters are strictly contractive). More precisely, one has the following result.

THEOREM 11. *Let there be given two matrix functions $X_{m-1,j}$ and $Y_{m-1,j}$ of class $L_2(H_{m-1,j})$, the inverses of which belong to $L_2(H)$. In addition, assume* (43) *and* (51) *to hold with m replaced by $m-1$. Then, for any $p \times p$ matrix $E_{m,j}$ subject to $\|E_{m,j}\| < 1$,* (55) *defines two functions $X_{m,j}$ and $Y_{m,j}$ in $L_2(H_{m,j})$, with inverses in $L_2(H)$, satisfying* (43) *and* (51). *Furthermore, for every point $(s, t) \in H_{m-1,j}$ the trigonometric moment of index (s, t) relative to the weight function $W_{m,j}$ coincides with the corresponding moment relative to $W_{m-1,j}$.*

Proof. The argument is quite similar to that given in Theorems 8 and 9. The only additional point is concerned with the inequalities (51). These clearly are hereditary (from $m-1$ to m) in view of

$$\begin{bmatrix} \tilde{Y}_{m,j}^{(0)} & X_{m,j}^{(j)} \\ \tilde{Y}_{m,j}^{(j)} & X_{m,j}^{(0)} \end{bmatrix} = \begin{bmatrix} \tilde{Y}_{m-1,j}^{(0)} & X_{m-1,j}^{(j)} \\ \tilde{Y}_{m-1,j}^{(j)} & X_{m-1,j}^{(0)} \end{bmatrix} Z_{m,j}, \tag{58}$$

owing to the fact that $Z_{m,j}$ is J-unitary. The role of (51) is to guarantee that $X_{m,j}(e^{i\theta}, 0)^{-1}$ and $Y_{m,j}(e^{i\theta}, 0)^{-1}$ belong to L_2^+. Details are left to the reader. □

6. Summability and strong convergence properties. This section has two main objects: first, to obtain strong stability and convergence properties of the minimizing polynomials $F_{m,j}$ and $G_{m,j}$, together with a computation method based on the Schur parameters; and second, to establish convergence of the minimizing trigonometric polynomials $F_{k,m,j}$ and $G_{k,m,j}$ to the stable functions $F_{m,j}$ and $G_{m,j}$ in a sense warranting asymptotic stability of the approximants. In addition to their theoretical interest, these topics have a definite significance from the application viewpoint [3], [6], [16].

Appropriate assumptions leading to the desired results turn out to be certain summability properties of the Fourier series of $\Gamma_j(\theta)$. We shall use the same Banach algebra techniques as in [7]. (These are mainly borrowed from Baxter [2]; see also Geronimo [9].) Let f be a real-valued function, defined over the integers, satisfying $f(-m) = f(m) \geqq 1$ and $f(m+n) \leqq f(m)f(n)$ for all $m, n \in \mathbb{Z}$. In addition, assume $\lim f(m)^{1/m} = 1$ for $m \to \infty$. (For example, $f(m) = (1+|m|)^c$ with a given $c \geqq 0$.) The *Banach algebra* B_f, with norm $|\cdot|_f$, is defined to consist of all integrable $p \times p$ matrix

functions $K(e^{i\theta}) \sim \sum K_s e^{is\theta}$ for which the numerical series

$$|K|_f = \sum_{s=-\infty}^{+\infty} f(s)\|K_s\| \tag{59}$$

is convergent. Next, given a subset S of the plane $\mathbb{Z} \times \mathbb{Z}$, with $(0, 0) \in S$, let us denote by $B_f(S)$ the class of two-variable functions $X(e^{i\theta}, e^{i\phi}) \sim \sum X^{(t)}(e^{i\theta}) e^{it\phi}$ belonging to $L_1(S)$, with $X^{(t)} \in B_f$ for all t, for which the series $|X|_f = \sum_t |X^{(t)}|_f$ converges. Clearly, $B_f(S)$ is a Banach space for the norm $|\cdot|_f$ thus defined. (Note that $B_f(S)$ is a Banach algebra when S is closed under addition.)

The following results show that the theory of the minimizing functions with interval support can be completely developed in the framework of the Banach algebras B_f.

THEOREM 12. *Assume that the trigonometric moments $\Delta_0, \Delta_1, \cdots, \Delta_j$ belong to B_f, for a given f. Then the normalized minimizing functions X_j, Y_j, $X_{m,j}$, $Y_{m,j}$ and their inverses belong to the class $B_f(H)$. In addition, the Schur parameters $E_{m,j}$ satisfy $\sum_{m=-\infty}^{+\infty} f(m)\|E_{m,j}\| < \infty$.*

Proof. Since Γ_j is nonsingular, the Wiener–Lévy theorem shows that the blocks of Γ_j^{-1} belong to B_f. Hence $A_{j,0}, A_{j,1}, \cdots, A_{j,j} \in B_f$, by (7), and thus $A_{j,0}^{-1} \in B_f$ (since $A_{j,0}$ is nonsingular). As a result, M_j and M_j^{-1} belong to B_f (see [10, p. 188]), so that (13) yields $X_j \in B_f(H_j)$. Similarly $Y_j \in B_f(H_j)$. Next, writing (18) in the form $\Omega_j = N_{j-1}^{-1} \sum_{t=0}^{j-1} \Delta_{j-t} X_{j-1,t}$ shows that Ω_j is of class B_f. This implies that the blocks of $U_{m,j}$ in (39) belong to B_f, for all $m \in \mathbb{Z}$ (see [7]), so that Theorem 7 leads to the conclusion $X_{m,j}, Y_{m,j} \in B_f(H_{m,j})$, which yields $X_{m,j}^{-1}, Y_{m,j}^{-1} \in B_f(H)$ in view of Theorem 8 (by application of the Wiener–Lévy theorem). Note that $X_j^{-1}, Y_j^{-1} \in B_f(H)$ follows similarly from Theorem 2. Finally, the assertion concerning the Schur parameters is taken from [7]. □

Remark. Note that $X_j(e^{i\theta}, w)$ is nonsingular in all disks $|w| \leqq R < R_0$, for some $R_0 > 1$. This yields a summability property of X_j^{-1} stronger than that quoted in Theorem 12, namely $X_j(e^{i\theta}, \mathrm{Re}^{i\phi})^{-1} \in B_f(H)$. Similar properties hold for Y_j, $X_{m,j}$ and $Y_{m,j}$.

THEOREM 13. *If $\Delta_0, \Delta_1, \cdots, \Delta_j \in B_f$, then the doubly infinite sequence of minimizing functions $F_{m,j}$ is convergent in the space $B_f(H_j)$; i.e., $|F_{m,j} - F_j|_f \to 0$ when $m \to +\infty$ and $|F_{m,j} - F_{j-1}|_f \to 0$ when $m \to -\infty$. Similar results hold for $G_{m,j}$.*

Proof. For a suitable normalization one has $U_{m,j} \to U_j$ when $m \to +\infty$ and $U_{m,j} \to I$ when $m \to -\infty$, in the sense of the norm $|\cdot|_f$ (see [7]). Hence the theorem follows from (40) and (20). □

THEOREM 14. *Assume $\sum_{m=-\infty}^{+\infty} f(m)\|E_{m,t}\| < \infty$ for $t = 0, 1, \cdots, j$. Then the functions $\Delta_0, \Delta_1, \cdots, \Delta_j$ belong to the Banach algebra B_f.*

Proof. The matrix version of Baxter's theorem [2] first yields $X_0, Y_0 \in B_f$ as a consequence of $\sum_{m=1}^{\infty} f(m)\|E_{m,0}\| < \infty$. Then it follows from Theorem 7 and the results of [7] that $X_{m,1}$, $Y_{m,1}$, X_1 and Y_1 belong to $B_f(H_1)$, with the convergence properties $|X_{m,1} - X_1|_f \to 0$ and $|Y_{m,1} - Y_1|_f \to 0$ for $m \to \infty$, owing to the assumption $\sum_{m=-\infty}^{+\infty} f(m)\|E_{m,1}\| < \infty$. Using this argument in an inductive manner one obtains $X_j \in B_f(H_j)$, which yields $X_j^{-1} \in B_f(H)$ in view of the Wiener–Lévy theorem. Hence the desired result follows from (22). □

In fact, it is possible to construct the minimizing functions with interval support by starting directly from the Schur parameters $E_{s,t}$. This is really interesting because the stability properties are then automatically satisfied (see Theorem 11). Let us briefly describe the computation scheme. We assume $\sum f(s)\|E_{s,t}\| < \infty$ for $0 \leqq t \leqq j$ as in Theorem 14. Define the J-unitary matrix $Z_{s,t}$ from the given Schur parameter $E_{s,t}$ as in (49). In addition, let the $2p \times 2p$ matrix $Z_{\infty,t} = wI \dotplus I$. Then $X_{m,j}(e^{i\theta}, w)$ and

$Y_{m,j}(e^{i\theta}, w)$ are formally given by

$$[\hat{Y}_{m,j} \quad X_{m,j}] = [\hat{Y}_{0,0} \quad X_{0,0}] \prod \{Z_{s,t} : (1,0) \leqq (s,t) \leqq (m,j)\}, \tag{60}$$

where the infinite product $\prod$ is performed in accordance with the extended inverse lexicographic ordering; i.e.,

$$\begin{aligned}(1,0) < (2,0) < \cdots < (\infty,0) < \cdots < (-2,1) < (-1,1) < (0,1) < (1,1)\\ < \cdots < (\infty,1) < \cdots < \cdots < (-2,j-1) < (-1,j-1) < (0,j-1)\\ < (1,j-1) < \cdots < (\infty,j-1) < \cdots < (m-2,j) < (m-1,j) < (m,j).\end{aligned}$$

Note that when the product in (60) is stopped at the index (∞, t) one obtains the matrix $[w\hat{Y}_t \quad X_t]$. The precise meaning of (60) is the following. For any integer α define both infinite products,

$$\begin{aligned} K^{+}_{\alpha,t} &= \lim_{k\to\infty} Z_{\alpha,t} Z_{\alpha+1,t} \cdots Z_{\alpha+k,t},\\ K^{-}_{\alpha,t} &= \lim_{k\to\infty} Z_{\alpha-k,t} Z_{\alpha-k+1,t} \cdots Z_{\alpha,t}. \end{aligned} \tag{61}$$

(Note that $K^{-}_{\alpha,t}$ coincides with $U_{\alpha,t}$ and $K^{+}_{\alpha,t}$ with the inverse of $V_{\alpha-1,t}$ defined as in (39) and (41).) From the matrix version of Baxter's theorem [2] it follows that the limits (61) exist in the space B_f. The product $\prod$ occurring in (60) has to be interpreted as

$$\prod = K^{+}_{1,0} W K^{-}_{0,1} K^{+}_{1,1} W K^{-}_{0,2} K^{+}_{1,2} W \cdots K^{-}_{0,j-1} K^{+}_{1,j-1} W K^{-}_{m,j}, \tag{62}$$

with $W = wI \dotplus I$. The algorithm resulting from (60) and (62) is easily understandable: truncating the sequence $(E_{m,t})$ for $m > k$ and $m < -k$, where k is a given positive integer, amounts to replacing (61) by the finite products $K^{+}_{\alpha,t} = Z_{\alpha,t} \cdots Z_{\alpha+k,t}$ and $K^{-}_{\alpha,t} = Z_{\alpha-k,t} \cdots Z_{\alpha,t}$. Then (60) and (62) produce the exact functions $X_{m,j}$, $Y_{m,j}$ in case $E_{m,t} = 0$ for $m \notin [-k, k]$ and approximate functions $X^{(k)}_{m,j}$, $Y^{(k)}_{m,j}$ in the general case; these approximants are stable and converge to the desired functions $X_{m,j}$, $Y_{m,j}$, in the sense of the norm $|\cdot|_f$, when k tends to infinity.

Let us conclude this section by studying another type of convergence, namely $F_{k,m,j} \to F_{m,j}$ and $G_{k,m,j} \to G_{m,j}$ for $k \to \infty$ (with fixed values of m and j). Remember that $F_{k,m,j}$ and $G_{k,m,j}$ are the minimizing trigonometric polynomials of support $H_{k,m,j}$ relative to the functionals Φ_r and Φ_l, respectively, and are thus easily computable from the coefficients $\Sigma_{s,t}$ by solving linear equations (see § 2).

THEOREM 15. *Assume* $\Delta_0, \Delta_1, \cdots, \Delta_j \in B_f$ *for a given function* f *such that* $\sum_{s=0}^{\infty} f(s)^{-2} < \infty$. *Let* g *be a positive even function satisfying*

$$\sup_{0 \leqq k < \infty} \left(\sum_{s=1}^{k} g(s)^2 \sum_{s=k}^{\infty} f(s)^{-2} \right) < \infty. \tag{63}$$

Then one has $|F_{k,m,j} - F_{m,j}|_g \to 0$ *and* $|G_{k,m,j} - G_{m,j}|_g \to 0$ *for* $k \to \infty$, *where* $|\cdot|_g$ *is defined as in* (59).

Proof. Let $Q_k \in L(H_{k,m,j})$ denote the truncation of $F_{m,j}$ to the support $H_{k,m,j}$. The first statement follows from juxtaposition of the following three inequalities:

$$|F_{k,m,j} - Q_k|_g \leqq (j+1)^{1/2} \left(\sum_{s=-k}^{k} g(s)^2 \right)^{1/2} \|F_{k,m,j} - F_{m,j}\|_2, \tag{64}$$

$$\|F_{k,m,j} - F_{m,j}\|_2 \leqq c \|Q_k - F_{m,j}\|_2, \tag{65}$$

$$\|Q_k - F_{m,j}\|_2 \leqq \varepsilon (2j+2)^{1/2} \Big(\sum_{s=k+1}^{\infty} f(s)^{-2} \Big)^{1/2}. \tag{66}$$

First, (64) is a straightforward consequence of Cauchy's inequality. As for (66), which is valid whenever $k \geqq k_0(\varepsilon)$, with ε denoting an arbitrarily small positive real number, it follows immediately from the property $f(s)\|F_{m,j}^{(s,t)}\| \leqq \varepsilon$ for $|s| \geqq k_0$ and $0 \leqq t \leqq j$ (see Theorem 12). Let us now establish (65). By definition, $\Phi(F_{k,m,j}) \leqq \Phi(Q)$ for any trigonometric polynomial Q of support $H_{k,m,j}$, hence $(F_{k,m,j} - F_{m,j}, F_{k,m,j} - F_{m,j}) \leqq (Q - F_{m,j}, Q - F_{m,j})$ by Theorem 4. Taking the trace of both members one obtains

$$\|X_j^{-1}(F_{k,m,j} - F_{m,j})\|_2 \leqq a\|Q - F_{m,j}\|_2, \tag{67}$$

by use of (26) and (27), for a suitable constant a. On the other hand, Poisson's inequality together with Theorem 2 yields

$$\|F_{k,m,j}(\cdot, w) - F_{m,j}(\cdot, w)\|_2 \leqq b\|X_j^{-1}(F_{k,m,j} - F_{m,j})\|_2 \tag{68}$$

for $|w| \leqq r < 1$, where b is a constant (depending on r). Owing to the fact that $F_{k,m,j}$ and $F_{m,j}$ are polynomials of fixed degree j in w, one readily deduces from (67) and (68) that the desired inequality (65) is satisfied for an appropriate constant c.

The end of the proof is immediate. From (64)–(66) one has $|F_{k,m,j} - Q_k|_g \to 0$ when $k \to \infty$, hence $|F_{k,m,j} - F_{m,j}|_g \to 0$ by definition of Q_k, owing to both properties $F_{m,j} \in B_f(H_{m,j})$ and $g(m) \leqq df(m)$ for a constant d. The second assertion of the theorem can be proved by the same method. □

COROLLARY 16. *Besides the conditions of Theorem* 15, *assume that g satisfies* $g(m+n) \leqq g(m)g(n)$, *for all* $m, n \in \mathbb{Z}$, *so that* B_g *is a Banach algebra. Then the minimizing trigonometric polynomials* $F_{k,m,j}(z, w)$ *are asymptotically stable, in the sense that they are nonsingular in the regions* $|z| = 1$, $|w| \leqq 1$ *and* $|z| \leqq 1$, $w = 0$, *provided* $k \geqq k_0(m, j)$. *In fact,* $F_{k,m,j}^{-1}$ *converges to* $F_{m,j}^{-1}$ *when* $k \to \infty$, *in the norm* $|\cdot|_g$. *Similar results hold for* $G_{k,m,j}(z, w)$.

Proof. It follows directly from Theorem 15 that $\det F_{k,m,j}(z, w)$ converges uniformly to $\det F_{m,j}(z, w)$ in the regions $|z| = 1$, $|w| \leqq 1$ and $|z| \leqq 1$, $w = 0$. Since $F_{m,j}$ is nonsingular in these regions (by Theorem 8), so is $F_{k,m,j}$ for all $k \geqq k_0(m, j)$. As a consequence, Theorem 15 implies convergence of $F_{k,m,j}^{-1}$ to $F_{m,j}^{-1}$ in the Banach algebra $B_g(H)$. □

Let us mention a simple application of Theorem 15 and Corollary 16. The condition (63) is clearly satisfied by the functions $f(m) = (1+|m|)^c$ and $g(m) = (1+|m|)^{c-1}$ with $c \geqq 1$. In particular, for the choice $c = 1$, Corollary 16 shows that $F_{k,m,j}$ is asymptotically stable and admits an inverse converging to $F_{m,j}^{-1}$ in the Wiener norm $|\cdot|_1$, provided $\Delta_0, \Delta_1, \cdots, \Delta_j \in B_f$ with $f(m) = 1 + |m|$.

7. Half-plane spectral factorization. Consider the Lebesgue decomposition $d\Sigma = W\, d\theta\, d\phi + d\Sigma_s$ for any nondecreasing $p \times p$ Hermitian-valued measure $\Sigma(\theta, \phi)$. By definition, W is the derivative of Σ, so that the measures $\int_0^\theta \int_0^\phi W(\alpha, \beta)\, d\alpha\, d\beta$ and $\Sigma_s(\theta, \phi)$ are the absolutely continuous part and the singular part of $\Sigma(\theta, \phi)$, respectively. Note that $W(\theta, \phi)$ is a nonnegative definite Hermitian-valued integrable function. Assume the logarithm of the determinant of $W(\theta, \phi)$ to be integrable. This is known to be a criterion for W to admit a *half-plane spectral factor* on the right, i.e., a $p \times p$ matrix-valued function $M(e^{i\theta}, e^{i\phi})$ of class $L_2(H)$, with $\det M(0, 0) \neq 0$, satisfying

$$W(\theta, \phi) = \tilde{M}(e^{i\theta}, e^{i\phi}) M(e^{i\theta}, e^{i\phi}). \tag{69}$$

Moreover, in this case there exists a *canonical* spectral factor, characterized by the

additional property

$$\log|\det M(0,0)| = \frac{1}{8\pi^2}\iint \log\det W(\theta,\phi)\,d\theta\,d\phi, \tag{70}$$

which uniquely determines M modulo premultiplication by a constant unitary matrix. (Note, for example, that (48) actually is such a spectral factorization.) Let us indicate the stability properties of the canonical spectral factor: $M(e^{i\theta}, w)$ is an outer function of w (almost everywhere in θ) and $M(z, 0)$ is an outer function of z. Under the same condition $\log\det W \in L_1$, there also exists a canonical spectral factor $N(e^{i\theta}, e^{i\phi}) \in L_2(H)$ on the left, characterized by $W(\theta,\phi) = N(e^{i\theta}, e^{i\phi})\tilde{N}(e^{i\theta}, e^{i\phi})$, together with (70) where N is substituted for M. The results above are due to Helson and Lowdenslager [11]. In addition, these authors have established an implicit version of Theorem 17 below. Note that the assumption $\log\det W \in L_1$ implies nondegeneracy of the inner products (1) with respect to any finite set S.

THEOREM 17. *Let $S_0 \subset S_1 \subset \cdots \subset S_n \subset \cdots$ be an ascending chain of finite subsets S_n of the upper half-plane H such that $\bigcup_{n=0}^{\infty} S_n = H$. Define $P_n(e^{i\theta}, e^{i\phi}) \in L(S_n)$ to be the minimizing trigonometric polynomial of support S_n relative to the functional Φ_r. Then, under the condition* $\log\det W \in L_1$, *one has*

$$\lim_{n\to\infty} \|I - \tilde{M}(0,0)MP_n\|_2 = 0, \tag{71}$$

where $M(e^{i\theta}, e^{i\phi})$ stands for the canonical spectral factor of $W(\theta,\phi)$ on the right. A similar result holds for the minimizing trigonometric polynomials $Q_n \in L(S_n)$ relative to Φ_l; one has $\|I - Q_n N\tilde{N}(0,0)\|_2 \to 0$, where N is the canonical spectral factor of W on the left.

Proof. We consider only the case $M(0,0) = I$, which is a simple matter of normalization. For any trigonometric polynomial Q of support H one has $\|I - MQ\|_2^2 = p - \operatorname{tr} Q(0,0) - \operatorname{tr} \tilde{Q}(0,0) + \|MQ\|_2^2 \leqq p + \operatorname{tr}\Phi(Q)$, in view of (2) and (69). On the other hand, it follows from [11, Thm. 12] that the infimum of $\operatorname{tr}(Q, Q)$ equals p when Q is subject to both $Q(0,0) \geqq 0$ and $\det Q(0,0) = 1$. As these conditions imply $\operatorname{tr} Q(0,0) \geqq p$, one obtains $\inf \operatorname{tr}\Phi(Q) \leqq -p$, by use of (2). Hence it appears that $p + \operatorname{tr}\Phi(P_n)$ tends monotonically to zero for $n \to \infty$, which yields the desired result. □

A rough interpretation of Theorem 17 is that P_n tends to the inverse of $\tilde{M}(0,0)M$ when $n \to \infty$. We shall now give a precise meaning to this property under the assumption of boundedness of both functions

$$u(\theta) = \frac{1}{2\pi}\int_0^{2\pi} \log\det W(\theta,\phi)\,d\phi, \qquad v(\theta) = \frac{1}{2\pi}\int_0^{2\pi} \operatorname{tr} W(\theta,\phi)\,d\phi. \tag{72}$$

LEMMA 18. *Let M be the canonical spectral factor of W on the right. If $u \in L_\infty$, then the scalar function* $\det M(e^{i\theta}, 0)$ *and its inverse belong to the class L_∞^+. If in addition $v \in L_\infty$, then the matrix function $M(e^{i\theta}, w)$ and its inverse are essentially bounded in θ, uniformly on every disk $|w| \leqq r < 1$. The same results hold for the canonical spectral factor on the left.*

Proof. By definition, $|\det M(e^{i\theta}, 0)|^2 = \exp u(\theta)$, so that both $\det M(e^{i\theta}, 0)$ and $\det M(e^{i\theta}, 0)^{-1}$ are bounded functions. Hence these functions belong to L_∞^+, in view of the fact that $\det M(z, 0)$ is an outer Hardy function. To prove the second statement we use the Poisson inequality, which yields

$$\tilde{M}(e^{i\theta}, w)M(e^{i\theta}, w) \leqq \frac{1+r}{1-r}\cdot\frac{1}{2\pi}\int_0^{2\pi} W(\theta,\phi)\,d\phi \tag{73}$$

in $|w| \leqq r$, as a consequence of (69). For a matrix function $X(e^{i\theta})$ define the norm $\|X\|_\infty = \text{ess sup } \|X(e^{i\theta})\|$, with $\|\cdot\|$ denoting the spectral norm. Then (72) and (73) yield

$$\|M(\cdot, w)\|_\infty^2 \leqq \frac{1+r}{1-r}\|v\|_\infty. \tag{74}$$

On the other hand, from the fact that $\det M(e^{i\theta}, w)$ is an outer function of w, one easily deduces the inequality

$$\log|\det M(e^{i\theta}, w)|^2 \geqq \frac{1+r}{1-r}\cdot\frac{1}{2\pi}\int_0^{2\pi} \log^- \det W(\theta, \phi)\, d\phi, \tag{75}$$

with $\log^- x = \min(0, \log x)$. Since $\log^- x \geqq \log x - px^{1/p}$, the property $p(\det W)^{1/p} \leqq \operatorname{tr} W$ applied to (75) leads to

$$|\det M(e^{i\theta}, w)|^2 \geqq \exp\left[-\frac{1+r}{1-r}(\|u\|_\infty + \|v\|_\infty)\right]. \tag{76}$$

Combining (74) and (76) and observing that $\|A\| \leqq \alpha$ and $|\det A| \geqq \beta$ imply $\|A^{-1}\| \leqq \beta^{-1}\alpha^{p-1}$, one obtains an inequality of the form $\|M^{-1}(\cdot, w)\|_\infty \leqq c(r)$. This concludes the proof. □

THEOREM 19. *Assume that u and v belong to L_∞. Then, in the situation of Theorem 17, the sequence of polynomials $P_n(\cdot, w)$ converges in the mean to the inverse of $\tilde{M}(0,0)M(\cdot, w)$, uniformly on every disk $|w| \leqq r < 1$. In the same sense, Q_n tends to $[N\tilde{N}(0,0)]^{-1}$ for $n \to \infty$.*

Proof. Applying the Poisson inequality to (71), one obtains

$$\|I - \tilde{M}(0,0)M(\cdot, w)P_n(\cdot, w)\|_2 \to 0$$

for $n \to \infty$, and hence the desired result by use of Lemma 18. □

Let us now restrict our attention to the situation (23) where $\Sigma(\theta, \phi)$ is absolutely continuous with respect to the variable θ. Note that $W(\theta, \phi)$ occurs as the derivative of the function $\Delta(\theta, \phi)$ with respect to ϕ. The condition $\log\det\Gamma_j \in L_1$ in (11) is always satisfied, for all j, as a consequence of $\log\det W \in L_1$. Indeed, one has (cf. [4])

$$\exp[(j+1)\operatorname{tr}\Delta_0(\theta)] \geqq \det\Gamma_j(\theta) \geqq |\det M(e^{i\theta}, 0)|^{2j+2}, \tag{77}$$

so that $\log|\det M(\cdot, 0)| \in L_1$ implies $\log\det\Gamma_j \in L_1$. Furthermore, in case $u \in L_\infty$ it follows from (77) and Lemma 18 that the condition $\det\Gamma_j^{-1} \in L_\infty$ in (24) is satisfied for all j. On the other hand, note that $\operatorname{tr}\Delta_0 \in L_\infty$ implies $v \in L_\infty$, since $0 \leqq v(\theta) \leqq \operatorname{tr}\Delta_0(\theta)$ by definition. It turns out that the results of Theorems 17 and 19 remain valid for arbitrary sets S_n of finite width (see Theorem 4).

THEOREM 20. *Assume both functions $\operatorname{tr}\Delta_0$ and u to be bounded. Let $S_0 \subset S_1 \subset \cdots \subset S_n \subset \cdots$ be an ascending chain of subsets of the upper half-plane H, such that $S_n \subset H_{j(n)}$ for all n (with H_j denoting the horizontal strip (28)), satisfying $\bigcup_{n=0}^\infty S_n = H$. Let $P_n(e^{i\theta}, w) \in L_2(S_n)$ denote the minimizing function of support S_n relative to the functional Φ_r. Then, for $n \to \infty$, one has both convergence properties* l.i.m. $MP_n = \tilde{M}(0,0)^{-1}$ *in the sense of the two-variable L_2-norm and* l.i.m. $P_n(\cdot, w) = [\tilde{M}(0,0)M(\cdot, w)]^{-1}$ *for the one-variable norm (uniformly on $|w| \leqq r < 1$). Similar results hold for the minimizing functions relative to the functional Φ_l.*

Proof. An argument similar to that given in Theorem 17 yields the first property. Then, application of Lemma 18 leads to the second assertion. □

From Theorems 19 and 20 one deduces, in particular, that the minimizing functions $F_{k,m,j}$, $F_{m,j}$ and F_j converge to the inverse of $\tilde{M}(0,0)M$ when k and j tend to

infinity. Note that these convergence properties hold true in the sense of the two-variable L_2-norm provided W^{-1} is a bounded function. The result $F_{k,m,j} \to [\tilde{M}(0,0)M]^{-1}$ is especially interesting because, under rather weak explicit conditions, it yields an inverse approximation of the canonical spectral factor of W by means of trigonometric polynomials which are both asymptotically stable (Corollary 16) and easily computable (Theorem 1). The idea of this approximation is due to Chang and Aggarwal [3].

Let us make a final remark concerning the Schur parameters $E_{m,j}$. It turns out that the integrability of $\log \det W$ is equivalent to the summability of the squared norms $\|E_{m,j}\|^2$ for (m, j) varying over the half-plane H. (The proof is similar to that given in [4] for the one-variable situation.) The interesting question of the (weighted) summability of the norms $\|E_{m,j}\|$ over H is more complicated and is left to further investigation.

REFERENCES

[1] V. M. ADAMJAN, D. Z. AROV AND M. G. KREIN, *Infinite Hankel block matrices and related extension problems*, Amer. Math. Soc. Transl., 111 (1978), pp. 133–156.

[2] G. BAXTER, *A convergence equivalence related to polynomials orthogonal on the unit circle*, Trans. Amer. Math. Soc., 99 (1961), pp. 471–487.

[3] H. CHANG AND J. K. AGGARWAL, *Design of two-dimensional semicausal recursive filters*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 1051–1059.

[4] P. DELSARTE, Y. GENIN AND Y. KAMP, *Orthogonal polynomial matrices on the unit circle*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 149–160.

[5] ———, *Schur parametrization of positive definite block-Toeplitz systems*, SIAM J. Appl. Math., 36 (1979), pp. 34–46.

[6] ———, *Half-plane Toeplitz systems*, IEEE Trans. Information Theory, IT-26 (1980), pp. 465–474.

[7] ———, *Generalized Schur representation of matrix-valued functions*, this Journal/this issue, pp. 94–107.

[8] M. P. EKSTROM AND J. W. WOODS, *Two-dimensional spectral factorization with applications in recursive digital filtering*, IEEE Trans. Acoust. Speech Signal Process., ASSP-24 (1976), pp. 115–128.

[9] J. S. GERONIMO, *Matrix orthogonal polynomials on the unit circle*, to be published.

[10] I. C. GOHBERG AND I. A. FEL'DMAN, *Convolution Equations and Projection Methods for Their Solution*, American Mathematical Society, Providence, RI, 1974.

[11] H. HELSON AND D. LOWDENSLAGER, *Prediction theory and Fourier series in several variables*, Acta Math., 99 (1958), pp. 165–202.

[12] I. I. HIRSCHMAN, JR., *Szegö polynomials on a compact group with ordered dual*, Canad. J. Math., 18 (1966), pp. 538–560.

[13] ———, *Szegö functions on a locally compact Abelian group with ordered dual*, Trans. Amer. Math. Soc., 121 (1966), pp. 133–159.

[14] ———, *Matrix-valued Toeplitz operators*, Duke Math. J., 34 (1967), pp. 403–415.

[15] ———, *Recent developments in the theory of finite Toeplitz operators*, in Advances in Probability and Related Topics, vol. 1, P. Ney, ed., Marcel Dekker, New York, 1971, pp. 104–167.

[16] T. L. MARZETTA, *A linear prediction approach to two-dimensional spectral factorization and spectral estimation*, Ph.D. dissertation, Dept. Electr. Eng. and Comp. Sci., Massachusetts Institute of Technology, Cambridge, MA, 1978.

[17] B. T. O'CONNOR AND T. S. HUANG, *Stability of general two-dimensional recursive digital filters*, IEEE Trans. Acoust. Speech Signal Process., ASSP-26 (1978), pp. 550–560.

[18] M. ROSENBERG, *The square-integrability of matrix-valued functions with respect to a non-negative Hermitian measure*, Duke Math. J., 31 (1964), pp. 291–298.

[19] N. WIENER AND P. MASANI, *The prediction theory of multivariate stochastic processes, Part* I, Acta Math. 98 (1957), pp. 111–150.

[20] A. S. WILLSKY, *Relationships between digital signal processing and control and estimation theory*, Proc. IEEE, 66 (1978), pp. 996–1017.

[21] D. C. YOULA AND N. N. KAZANJIAN, *Bauer-type factorization of positive matrices and the theory of matrix polynomials orthogonal on the unit circle*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 57–69.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 3, September 1981

0196-5212/81/0203-0001 $01.00/0

EQUILIBRIA ON A CONGESTED TRANSPORTATION NETWORK*

H. Z. AASHTIANI† AND T. L. MAGNANTI†

Abstract. Network equilibrium models arise in applied contexts as varied as urban transportation, energy distribution, spatially separated economic markets, electrical networks and water resource planning. In this paper, we propose and study an equilibrium model for one of these applications, namely for predicting traffic flow on a congested transportation network. The model is quite similar to those that arise in most contexts of network equilibria, however, and the methods that we use are applicable in these other settings as well.

Our transportation model includes such features as (i) multiple modes of transit, (ii) link interactions and their effect on congestion, (iii) limited choices (or perceptions) of paths for flow between any origin-destination pair, (iv) generalized cost or disutility for travel, and (v) demand relationships for travel between origin-destination pairs that depend upon the travel time (cost) between all other origin-destination pairs. Using Brouwer's fixed-point theorem, we establish existence of an equilibrium solution to the model. By imposing monotonicity conditions on the delay and demand functions, we also show that travel times (costs) are unique and, in certain instances, that link flows are unique.

1. Introduction. Network analysis draws its origins from several sources. Prominent among these is the study of passive electrical networks, particularly the prediction of a network's utilization when it is loaded with prescribed voltages and impedances. With given voltages applied to an electrical network, what is the resulting flow? More recently, similar types of predictive questions have been posed in social and economic contexts. In transportation, travelers' demands for transportation services function, like voltages, as forces that generate network flow which, in this instance, are trips to be made between origin and destination points in the network. In this setting, travel time, travel cost, and other disutility measures replace electrical resistance as the impedance to flow. In economics, price differentials between spatially separated markets act like voltages as forces for generating commodity flow; transportation costs between the markets act as resistance to commodity movement. In each of these applications, the equilibration of forces and impedances has served as a model for predicting flow on the network. The nature of the specific equilibrium model depends upon the behavioral assumption, such as Ohm's law, profit maximization or cost minimization, that relate the forces, impedances and network flow.

The advent of robust theories for constrained optimization has precipitated an attractive and common approach for studying network equilibrium problems, namely to view the equilibrium model as the Lagrange multiplier conditions or, more generally, the Karush–Kuhn–Tucker optimality conditions of well-conceived auxiliary optimization problems. For example, one might minimize power loss instead of finding an equilibrium on an electrical network directly. Making this association permits the powerful and flexible solution techniques of constrained optimization to be used to compute an equilibrium and, moreover, permits optimization theory to serve as the methodology base to study questions such as existence and uniqueness of equilibrium solutions. On the other hand, the equivalent optimization approach limits the richness of equilibrium modeling by restricting the problem assumptions to those for which the

* Received by the editors April 1, 1980, and in final form January 7, 1981. The material in this paper was presented at the TIMS XXIV International Meeting, Hawaii, June 1979. This research was supported by the Transportation Advanced Research Program of the U.S. Department of Transportation under contract DOT-TSC-1058 and by the National Science Foundation under grant 79-26625-ECS.

† Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

equilibrium conditions can be interpreted as optimality conditions for an associated optimization problem.

In this paper, we study a class of network equilibrium problems with no known equivalent optimization problem. Although the approach that we take might apply to a variety of different network equilibrium applications, we restrict our discussion to transportation planning. In the next section, we propose a general model for network equilibrium of an urban transportation system. The model includes such features as (i) multiple (and interacting) modes of transit, (ii) link interactions and their effect on congestion, (iii) limited choices (or perceptions) of paths for flow between any origin-destination pair, (iv) generalized cost or disutility for travel on any path that depends upon the flow pattern on the entire transportation network and (v) demand relationships for travel between origin-destination pairs that depend upon the travel time (cost) between *all* other origin-destination pairs. With the exception of (iii), any one of these modeling features invalidates the assumptions that are typically made when showing that the transportation equilibrium problem can be converted to an equivalent optimization model.

After stating this model and discussing some of its applications and specializations, we show that only very mild restrictions need be imposed upon the problem data, restrictions that we would expect to be met almost always in practice, to insure that an equilibrium solution exists. We also establish conditions that will insure that an equilibrium solution is unique. To establish these results, we formulate the equilibrium model as an equivalent nonlinear complementarity problem. Then we use Brouwer's fixed-point theorem to establish existence and nonlinear complementarity results to establish uniqueness.

2. Background. The genesis of transportation equilibrium modeling was a behavioral assumption, known as Wardrop's user traffic equilibrium law, first proposed in 1952 by the traffic engineer J. G. Wardrop [52], namely:

> At equilibrium, for each origin-destination pair the travel times on all the routes actually used are equal, and less than the travel times on all nonused routes.

This principle has spawned a great deal of research by transportation engineers, economists and operations researchers aimed at enhancing the scope and realism of Wardrop's model, at developing algorithms to compute an equilibrium, and at applying the equilibrium model in practice to predict traffic flow patterns. Modeling efforts and methodological advancements have evolved to the point that one version of the equilibrium model now forms part of the Urban Mass Transit Authority's transportation planning system [51].

Since 1952, a large number of algorithms have been developed for the traffic assignment problem. Most of the earlier techniques were heuristics and usually did not consider congestion effects or any formal concept of an equilibrium [39], [40], [53], [19]. The goal of these approaches was to assign flow between different paths so that the paths have almost equal travel time. The next generation of heuristics, as embodied by the "capacity restrained" technique [12], [28], [29], [48], attempted to account for capacity of the system. Later techniques [30], [38], [39] loaded the system incrementally, attempting to approximate an equilibrium solution.

The mathematical programming approach to traffic equilibrium originated in 1956 when Beckman, McGuire and Winsten [7] formulated a version of the equilibrium problem as the optimality conditions of an equivalent optimization problem.[1] They

[1] Samuelson had earlier proposed a similar transformation in the context of spatially separated economic markets.

assumed:

(1) a single mode of transit (private vehicle traffic has been the primary application since then);
(2) that the demand function $D_i(u_i)$ between every origin-destination pair i depends only upon the impedance or shortest travel time u_i between that origin-destination pair;
(3) that the delay functions for the links are separable; that is, the delay $t_a(v_a)$ for each link "a" depends only upon the total volume of traffic flow v_a on that link.

Since then several researchers have proposed algorithms for solving the equivalent optimization problem (Bruynooghe, Gibert and Sakarovitch [11], Bertsekas [8], Bertsekas and Gafni [9], Dafermos [13]–[16], Dembo and Klincewicz [18], Leventhal, Nemhauser and Trotter [36], Leblanc [34], [35], Nguyen [41]–[45], Golden [26], and Florian and Nguyen [23]–[25]).

There are a number of ways to enrich the modeling assumptions (1)–(3). Modeling multi-modal (for example, private vehicle and a public transit mode) and multi-class (for example, high vs. low income) traffic equilibrium would be extensions with great practical relevance. Incorporating demand functions for an O-D pair that depend upon impedance between other O-D pairs would permit destination choice to be modeled more realistically than in models based upon (1)–(3). For example, the distribution of trips from a residential district to two shopping centers depends, in part, upon the travel time to both centers. Residential home selection might be modeled as an origin choice version of this extension. Another extension would be to let delay on a link depend on volume flow on other links. This latter extension permits modeling of traffic equilibrium with two-way traffic in one link, traffic equilibrium with right and left turn penalties, and the like.

Some attempts have been made to generalize the equivalent optimization approach to traffic equilibrium to incorporate these modeling extensions. Dafermos [13], [15] has considered multiple classes of users, and Florian [22] and Abdulaal and Leblanc [4] have considered the multi-modal problem. In addition, the equivalent optimization problem has been used to prove existence and uniqueness of an equilibrium for certain specializations of the general model (Dafermos [13], [15], Florian and Nguyen [23] and Steenbrink [49]). Nevertheless, the optimization based approach is limited since the assumptions required to insure an equivalent convex optimization problem are generally too severe to be applicable in practice for modeling the type of extensions to assumptions (1)–(3) suggested above. The approach adopted in this paper originates with Aashtiani [1] who formulated an extended equilibrium model and studied existence of a solution by viewing the model as a nonlinear complementarity problem. In [2] he elaborates on this approach and proposes a computational scheme for solving for an extended equilibrium. Independently, Kuhn [27] devised a fixed-point method, equipped with a special pivoting scheme, to solve equilibrium problems with fixed demands and with separable link delay functions. Asmuth [6] has proposed an additive model similar to the one discussed in this paper that includes point-to-set delay functions and demand functions. He has also studied existence and uniqueness, existence being a consequence of a constructive fixed-point algorithm. The proof of existence given in this paper, which is adopted from Aashtiani and Magnanti [3], is shorter than these earlier proofs and relies on the classical fixed-point theorem of Brouwer.

In related developments, Dafermos [14], [15], by assuming differentiability and strong monotonicity of the link delay function, has recently used the theory of variational equalities to establish the existence of a traffic equilibrium and to devise an

algorithm for computing an equilibrium. Ahn [5] has used similar methods to study equilibrium for spatially separated markets arising in energy planning. Recently, Braess and Koch [10] and Smith [47] have used a proof different from that given in this paper, but also based upon Brouwer's fixed-point theorem, to establish existence of an equilibrium for a special version of the model that we study here; they assume that the demand is fixed independent of the network congestion and that the cost on any path is the sum of costs on arcs in that path. Braess and Koch also impose a monotonicity assumption on the arc costs.

3. Traffic equilibrium model. The equilibrium model is defined on a transportation network $[N, A]$ with nodes N, directed arcs A, and with a given set I of origin-destination (O-D) node pairs. Nodes represent centroids of population, business districts, street intersections and the like, and arcs model streets and arteries or might be introduced to model connections (and wait time) between legs of a trip, between modes, or between streets at an intersection. The model is formulated as:

$$
\begin{aligned}
&\text{(a)} && (T_p(h)-u_i)h_p = 0 && \text{for all } p \in P_i \text{ and } i \in I,\\
&\text{(b)} && T_p(h)-u_i \geqq 0 && \text{for all } p \in P_i \text{ and } i \in I,\\
(3.1)\quad &\text{(c)} && \sum_{p \in P_i} h_p - D_i(u) = 0 && \text{for all } i \in I,\\
&\text{(d)} && h \geqq 0, &&\\
&\text{(e)} && u \geqq 0. &&
\end{aligned}
$$

In this formulation:

- I is the set of O-D pairs.
- P_i is the set of "available" paths for flow for O-D pair i (which might, but need not, be all paths joining the O-D pair).
- h_p is the flow on path p.
- h is the vector of $\{h_p\}$ with dimension $n_1 = \sum_{i \in I} |P_i|$ equal to the total number of O-D pairs and path combinations.
- u_i is an accessibility variable, shortest travel time (or generalized cost) for O-D pair i.
- u is the vector of $\{u_i\}$ with dimension $n_2 = |I|$.
- $D_i(u)$ is the demand function for O-D pair i; $D_i : R_+^{n_2} \to R_+^1$.
- $T_p(h)$ is the delay time, or general disutility, function for path p; $T(h) : R_+^{n_1} \to R_+^1$.

We also let $P = \cup\{P_i : i \in I\}$ denote the set of all "available" paths in the network and assume that the network is strongly connected; i.e., for any O-D pair $i \in I$ there is at least one path joining the origin to the destination (i.e., $|P_i| \geqq 1$).

The first two equations in (3.1) model Wardrop's traffic equilibrium law requiring that for any O-D pair i, the travel time (generalized travel time) for all paths $p \in P_i$, with positive flow $h_p > 0$, is the same and equal to u_i, which is less than or equal to the travel time for any path with zero flow. Equation (3.1c) requires that the total flow among different paths between any O-D pair i equal the total demand $D_i(u)$, which in turn depends upon the congestion in the network through the shortest path variable u. Conditions (3.1d) and (3.1e) state that both flow on paths and minimum travel times should be nonnegative.

An important special case of the equilibrium problem (3.1) is an *additive model* in which

$$T_p(h) = \sum_{a \in A} \delta_{ap} t_a^i(h) \quad \text{for all } p \in P_i \text{ and } i \in I, \tag{3.2}$$

where

$$\delta_{ap} = \begin{cases} 1 & \text{if link } a \text{ is in path } p, \\ 0 & \text{otherwise,} \end{cases}$$

and $t_a^i(h)$ is the delay function for arc a and O-D pair i,

$$t_a^i : R_+^{n_1} \to R_+^1.$$

That is, the delay time on path p is the sum of the delays of the arcs in that path. More compactly, $T_p(h) = \Delta^T t^i(h)$, where $\Delta = (\delta_{ap})$ is the arc-path incidence matrix for the network and $t^i(h) = (t_a^i(h))$ is the vector of arc delay functions for O-D pair i.

Several features of the equilibrium model are worth noting. In a large transportation network, users generally will not perceive, or choose from, all possible paths joining every origin-destination pair. If we identify the paths P_i available for flow between O-D pair i as the available set of routes from which the user chooses, the equilibrium conditions model this type of limited route choice.[2] In addition, since the path disutility functions $T_p(h)$ are arbitrary and depend upon the full vector h of path flows, the model can account for path interactions, as at intersections, and the generalized costs $T_p(h)$ can, in principle, incorporate a variety of attributes that are relevant to route selection such as travel time, travel costs, and route attractiveness. To the best of our knowledge, no previous existence proof of traffic equilibrium incorporates both of these modeling features.

The equilibrium model (2.1) is more general than first appearance might indicate. A judicious choice of network structure permits the formulation to model a wide range of equilibrium applications including multi-modal transit, multiple classes of users and destination or origin choice. To model multi-modal situations, we might conceptualize an extended network with a distinct component for each mode of transit. (Dafermos [13] and Sheffi [46] adopt this approach as well.) The component networks might be identical copies of the underlying physical transportation network, as when autos and buses share a common street network. Since the delay $T_p(h)$ for paths on the automobile component network depends upon the full vector h of path flows, the delay function can account for congestion added by buses sharing common links. Note, however, that the networks for each mode need not be the same. Consequently, bus routes might be fixed and subway links might be distinct from those of other modes.

The model also provides flexibility in modeling demand. Suppose, for instance, that O-D pairs i and j in the extended network introduced above correspond to the same physical origin and destination points but different modes of transit. If we introduce a source node s and terminal node t connected, respectively, to the origin and destination points of O-D pairs i and j, then a demand function $D_{st}(u)$ would model

[2] Several authors (e.g., Asmuth [6], Dafermos [14], [16] and Smith [47]) formulate the traffic equilibrium problem in terms of arc flows. The path flow formulation with limited path choice appears to be more general. If A_i is the union of the arcs continued on the paths in P_i, then the arc formulation implies that any path with arcs in A_i and joining O-D pair i belongs to P_i. In formulation (3.1), P_i is an arbitrary collection of paths joining O-D pair i, thereby permitting more flexibility in modeling user's perception of "available" paths.

total trips between the origin and destination points as a function of network congestion. The equilibrium model would distribute these trips between the two modes to equalize the disutility $T_p(h)$ on all flow carrying paths by both modes. As an alternative, the modeler could prescribe the nature of modal split by introducing demand functions such as the well-known logit model:

$$D_i(u) = d\frac{e^{\theta u_i + A_i}}{e^{\theta u_i + A_i} + e^{\theta u_j + A_j}}, \qquad D_j(u) = d - D_i(u),$$

which would distribute the total number of trips d between the two modes i and j depending upon delay times u_i and u_j by the two modes and the given negative constant θ and nonnegative constants A_i and A_j.

As Dial [20] has noted, a generalized version of the logit model permits destination choice and modal split to be made simultaneously. If $i = pqm$ denotes an origin destination pair p–q distinguished by transit mode m, the model is of the form

$$D_{pqm}(u) = d_p\frac{r_q e^{\theta u_{pqm}}}{\sum_{q'}\sum_{m'} r_{q'} e^{\theta u_{pq'm}}},$$

where d_p is the total number of trips generated at origin p to be sent to the destination q', and $r_{q'}$ is an index of attraction for destination q'.

4. Equivalent nonlinear complementarity problem. Let $F(x) = (F_1(x), \cdots, F_n(x))$ be a vector-valued function from an n-dimensional space R^n into itself. The well-known nonlinear complementarity problem of mathematical programming is to find a vector x that satisfies the following system:

$$x \cdot F(x) = 0, \quad F(x) \geqq 0, \quad x \geqq 0. \tag{4.1}$$

This problem has wide ranging applications. Karamardian [31], [32] illustrates several examples. For instance, the primal-dual optimality conditions of linear and quadratic programming and the Kuhn–Tucker conditions for certain other nonlinear programming problems can be cast in this form.

In this section we show that the traffic equilibrium problem (3.1) can be formulated as a complementarity problem. By definition, equations (3.1a), (3.1b), and (3.1d) are complementary in nature. To show that the remaining equations can be expressed in a complementarity form requires some mild assumptions that we would expect to be met always in practice.

First some simplification in the formulation helps to clarify our discussion. Let $x = (h, u) \in R^n$, where $n = n_1 + n_2$, and furthermore, let

$$f_p(x) = T_p(h) - u_i \quad \text{for all } p \in P_i \text{ and } i \in I,$$

and

$$g_i(x) = \sum_{p \in P_i} h_p - D_i(u) \quad \text{for all } i \in I.$$

Also, let

$$F(x) = (f_p(x) \text{ for all } p \in P_i \text{ and } i \in I,\ g_i(x) \text{ for all } i \in I) \in R^n.$$

Then F is a vector-valued function from an n-dimensional space R^n into itself. Now

consider the following nonlinear complementarity system:

$$
\begin{aligned}
f_p(x)h_p &= 0 \qquad \text{for all } p \in P_i \text{ and } i \in I,\\
f_p(x) &\geqq 0 \qquad \text{for all } p \in P_i \text{ and } i \in I,\\
g_i(x)u_i &= 0 \qquad \text{for all } i \in I,\\
g_i(x) &\geqq 0 \qquad \text{for all } i \in I,\\
x &\geqq 0,
\end{aligned}
\tag{4.2}
$$

which is a specialization of (4.1).

Since any solution $\bar{x} = (\bar{h}, \bar{u})$ to the traffic equilibrium problem satisfies $g_i(\bar{x}) = 0$ for all $i \in I$, the solution $\bar{x}$ solves the nonlinear complementarity problem (4.2) as well, independent of the nature of the delay functions $T_p(h)$ and the demand functions $D_i(u)$. The following result establishes a partial converse.

PROPOSITION 4.1. *Suppose, for all $p \in P$, that $T_p : R_+^{n_1} \to R_+^1$ is a positive function. Also, suppose, for all $i \in I$, that $D_i : R_+^{n_2} \to R_+^1$ is a nonnegative function. Then the traffic equilibrium system* (3.1) *is equivalent to the nonlinear complementarity system* (4.2).

Proof. In light of our comment preceding the proposition, it is sufficient to show that any solution to (4.2) is a solution to (3.1). Suppose to the contrary that there is an $x = (h, u)$ satisfying (4.2), but that $g_i(x) = \sum_{p \in P_i} h_p - D_i(u) > 0$ for some $i \in I$. Then $g_i(x)u_i = 0$ implies that $u_i = 0$. Also, since D_i is nonnegative $\sum_{p \in P_i} h_p > D_i(u) \geqq 0$, which implies that $h_p > 0$ for some $p \in P_i$. But, for this particular p, the equation $f_p(x)h_p = 0$ implies that

$$f_p(x) = T_p(h) - u_i = 0 \quad \text{or} \quad T_p(h) = u_i.$$

But since $u_i = 0$, $T_p(h) = 0$, which contradicts the assumption $T_p(h) > 0$. □

When the traffic equilibrium problem is additive, $T_p(h) = \sum_{a \in A} \delta_{ap} t_a(h)$, $T_p(h)$ is positive whenever the arc delay functions are positive, or more generally, whenever the arc delay functions are nonnegative and at least one is positive on an arc a in path p.[3]

PROPOSITION 4.2. *Suppose, for all $a \in A$, that $t_a : R_+^{n_1} \to R_+^1$ is a positive function. Also, suppose, for all $i \in I$, that $D_i : R_+^{n_2} \to R_+^1$ is a nonnegative function. Then the additive traffic equilibrium system* (3.1) *and* (3.2) *is equivalent to the nonlinear complementarity system* (4.2).

Neither of the previous two propositions is valid if either the assumption that each demand function $D_i(u)$ is nonnegative or the assumption that each delay function $T_p(h)$ is positive is eliminated. See Aashtiani [2] for examples.

5. Existence. Rather extensive theory (see, for example, Karamardian [31] and Kojima [33]) provides necessary conditions that assure the existence of a solution to the nonlinear complementarity problem. Unfortunately, most of the conditions are too strong to be applied directly to the traffic equilibrium problem. To illustrate this situation and at the same time introduce concepts that will be useful in § 6 when we discuss uniqueness of solutions, we introduce a prototype of this theory by considering results due to Karamardian. First, we require some definitions.

[3] Notice that we have suppressed explicit dependence of the arc delay functions $t_a^i(h)$ on the origin-destination pair i since the generality of the equilibrium problem (3.1) permits us, at least conceptually, to duplicate the network, as indicated in the previous section, so that each arc carries the flow for a single O-D pair.

DEFINITION 5.1. Let $F: D \to R^n$, $D \subset R^n$. The function F is *monotone* on D if, for every pair $x \in D$ and $y \in D$,

$$(x-y)(F(x)-F(y)) \geqq 0.$$

F is *strictly monotone* on D if, for every pair $x \in D$, $y \in D$ with $x \neq y$,

$$(x-y)(F(x)-F(y)) > 0.$$

F is said to be *strongly monotone* on D if there is a scalar $k > 0$ such that, for every pair $x \in D$, $y \in D$,

$$(x-y)(F(x)-F(y)) \geqq k|x-y|^2,$$

where $|\cdot|$ denotes the usual Euclidean norm.

THEOREM 5.1 (Karamardian [31]). *If $F: R^n_+ \to R^n$ is continuous and strongly monotone on R^n_+, then the nonlinear complementarity system* (4.1) *has a unique solution.*

THEOREM 5.2 (Karamardian [31]). *If $F: R^n_+ \to R^n$ is strictly monotone on R^n_+, then the nonlinear complementarity system* (4.1) *has at most one solution.*

Notice that for traffic equilibrium problems, these theorems require that $F(x) = (\sum_{a \in A} T_p(h) - u_i$ for all $p \in P_i$ and $i \in I$, $\sum_{p \in P_i} h_p - D_i(u)$ for all $i \in I)$ and necessarily $T_p(h)$ be strictly or strongly monotone in terms of *path flows*. In most instances, this condition is not applicable; usually, the delay functions T_p depend upon arc flows each of which depends upon the sum of the flows on different paths. In these situations, whenever $x = (h, u)$ and $y = (h', u)$ correspond to two path flows h and h' that give rise to identical arc flows, $T_p(h) = T_p(h')$ and $\sum_{p \in P_i} h_p = \sum_{p \in P_i} h'_p$ for all $i \in I$. Consequently, $F(x) = F(y)$ and $(x-y)[F(x)-F(y)] = 0$, so that neither strict nor strong monotonicity applies.

Generally, however, for transportation applications the delay functions $T_p(h)$ are monotone, and frequently even strictly monotone, in terms of *link volumes*. Later we use this property to show the uniqueness of the solution in terms of link flows. In Theorem 5.3 to follow, though, we show that no monotonicity assumption is required for the existence of the solution.

To establish this result we use a well-known [50] transformation that permits us to convert the nonlinear complementarity problem and, in particular, the nonlinear complementarity version (4.2) of the traffic equilibrium problem into a Brouwer fixed-point problem. Let us define $\phi: R^n \to R^n$ by defining its component functions ϕ_i for $i = 1, 2, \cdots, n$ as

$$\phi_i(x) = [x_i - F_i(x)]^+,$$

where $[y]^+$ denotes $\max\{0, y\}$. Then $\bar{x}$ is a fixed point to ϕ; i.e., $\bar{x} = \phi(\bar{x})$ if and only if $\bar{x}$ solves the nonlinear complementarity problem $x \geqq 0$, $F(x) \geqq 0$, and $xF(x) = 0$.

This equivalence shows that we can, in principle, study any nonlinear complementarity problem by invoking fixed-point theory. Note that we cannot use Brouwer's fixed-point theorem directly, though, because the mapping $\phi(x)$ defined on R^n_+ need not map any compact set into itself. Consequently, we will restrict the domain of ϕ to some large cube C. To apply the theorem, we must be assured that ϕ maps C into itself, which we accomplish by redefining $\phi(x)$ for any $x \in C$ if it lies outside of C by projecting $\phi(x)$ onto C. By Brouwer's fixed-point theorem the modified map ϕ' has a fixed point. We must show that it has no *false* fixed points, though—that is, no point $\bar{x}$ contained on the boundary of C with the property that $\phi(\bar{x}) \notin C$ but the projection $\phi'(\bar{x})$ of $\phi(\bar{x})$ on C satisfies $\phi'(\bar{x}) = \bar{x}$. The essence of the following equilibrium proof is that ϕ' as derived

from the complementarity version (4.2) of the traffic equilibrium problem admits no false fixed points.

THEOREM 5.3. *Suppose* (N, A) *is a strongly connected network. Suppose that* $T_p : R_+^{n_1} \to R^1$ *is a nonnegative continuous function for all* $p \in P$. *Also suppose that for all* $i \in I$, $D_i : R_+^{n_2} \to R^1$ *is a continuous function that is bounded from above. Then the nonlinear complementarity system* (4.2) *has a solution.*

Proof. Let $F_i(h) = \sum_{p \in P_i} h_p$ denote the flow between O-D pair i and let e and $\hat{e}$ denote vectors of ones with $|P|$ and $|I|$ components. We must show that the following complementarity problem has a solution:

$$\left.\begin{aligned} h_p[T_p(h) - u_i] &= 0 \\ u_i[F_i(h) - D_i(u)] &= 0 \\ T_p(h) - u_i &\geqq 0 \\ F_i(h) - D_i(u) &\geqq 0 \\ u_i \geqq 0,\ h_p &\geqq 0 \end{aligned}\right\} \quad \text{for all } i \in I \text{ and all } p \in P_i.$$

Let $K_1 > 0$ satisfy

$$K_1 > \max_i \max_{u \geqq 0} D_i(u)$$

and let $K_2 \geqq K_1$ satisfy

$$K_2 > \max_{p \in P} \max_{0 \leqq h \leqq K_1 e} T_p(h).$$

K_1 exists because of the hypothesis that each $D_i(u)$ is bounded, and K_2 exists because each $T_p(h)$ is continuous.

Define the continuous mapping ϕ of the cube $\{0 \leqq h \leqq K_1 e,\ 0 \leqq u \leqq K_2 \hat{e}\}$ into itself by

$$\phi_p(h, u) = \min\{K_1, [h_p + u_i - T_p(h)]^+\} \quad \text{for all } p \in P_i \text{ and all } i \in I$$

and

$$\phi_i(h, u) = \min\{K_2, [u_i + D_i(u) - F_i(h)]^+\} \quad \text{for all } i \in I.$$

By Brouwer's fixed-point theorem this mapping has a fixed point $(\hat{h}, \hat{u})$; that is, $\hat{h}_p = \phi_p(\hat{h}, \hat{u})$ and $\hat{u}_i = \phi_i(\hat{h}, \hat{u})$ for all $i \in I$ and all $p \in P$. We show that this fixed point solves the complementarity problem by showing that, for all $p \in P_i$ and $i \in I$,

$$(*) \qquad \hat{h}_p = [\hat{h}_p + \hat{u}_i - T_p(\hat{h})]^+, \qquad \hat{u}_i = [\hat{u}_i + D_i(\hat{u}) - F_i(\hat{h})]^+.$$

First note that $\hat{u}_i < K_2$ for all $i \in I$, for if some $\hat{u}_i = K_2$ then, for any $p \in P_i$, $\hat{h}_p + \hat{u}_i - T_p(\hat{h}) > \hat{h}_p$ by the definition of K_2, which implies from $\hat{h}_p = \phi_p(\hat{h}, \hat{u})$ that $\hat{h}_p = K_1$. But then the definition of K_1 implies that $D_i(\hat{u}) < F_i(\hat{h})$, so that $[\hat{u}_i + D_i(\hat{u}) - F_i(\hat{h})] < \hat{u}_i$; therefore $\hat{u}_i$ must equal 0 in order that $\hat{u}_i = \phi_i(\hat{h}, \hat{u})$, contradicting $\hat{u}_i = K_2$.

Next note that if $h_p = K_1$ for some $i \in I$ and $p \in P_i$, then $D_i(\hat{u}) < F_i(\hat{h})$ by definition of K_1, which implies as above that $\hat{u}_i = 0$. By the nonnegativity of T_p, $[\hat{h}_p + \hat{u}_i - T_p(\hat{h})] \leqq \hat{h}_p$, with a strict inequality if $T_p(\hat{h}) > 0$. Consequently, in order that $\hat{h}_p = K_1 > 0$ equal $\phi_p(\hat{h}, \hat{u})$, $T_p(\hat{h})$ must equal 0 and thus $\hat{h}_p = [\hat{h}_p + \hat{u}_i - T_p(\hat{h})]^+$.

We have now established the expressions (*) which imply, if we consider the cases $\hat{h}_p > 0$ or $\hat{h}_p = 0$ and $\hat{u}_i > 0$ or $\hat{u}_i = 0$, that $(\hat{h}, \hat{u})$ solves the complementarity problem (4.2). □

As a consequence of Theorem 5.3 and Proposition 4.1 we have the following result.

THEOREM 5.4. (Existence). *Suppose* (N, A) *is a strongly connected network. Suppose that* $T_p : R_+^{n_1} \to R_+^1$ *is a positive continuous function for all* $p \in P$. *Also suppose that for all* $i \in I$, $D_i : R_+^{n_2} \to R_+^1$ *is a nonnegative continuous function that is bounded from above. Then the traffic-equilibrium system* (3.1) *has a solution.*

An important version of this theorem is its specialization for additive traffic equilibrium .

THEOREM 5.5. (Existence). *Suppose* (N, A) *is a strongly connected network. Suppose that* $t_a : R_+^{n_1} \to R_+^1$ *is a positive continuous function for all* $a \in A$. *Also suppose that for all* $i \in I$, $D_i : R_+^{n_2} \to R_+^1$ *is a nonnegative continuous function that is bounded from above. Then the additive traffic equilibrium system* (3.1) *and* (3.2) *has a solution.*

Proof. Since every t_a is positive and continuous, so is $T_p(h) = \sum_{a \in A} \delta_{ap} t_a(h)$ and, consequently, Theorem 5.4 applies. □

Asmuth [6] has suggested what appears to be a stronger version of Theorem 5.5 by not requiring that the demand functions $D_i(u)$ be bounded. To see the relevance of this result, suppose that $D_i(u)$ denotes the number of trips to be made between a particular origin-destination pair by automobiles. One possibility for modeling this situation is a Cobb–Douglas product form demand model given by

$$D_i(u) = A \frac{(u_j)^\beta}{(u_i)^\alpha},$$

where A is a given constant, $\alpha \geqq 0$ is a "direct elasticity" and $\beta \geqq 0$ is a "cross elasticity". In this model, u_i denotes the travel time between the O-D pair by auto and u_j denotes the travel time by an alternate mode such as bus. Note that $D_i(u)$ is not bounded unless we require $u_i \geqq \varepsilon$ for some, possibly small, number $\varepsilon > 0$.

The next result shows that Theorem 5.4 can be modified easily to include settings of this nature.

THEOREM 5.6. (Existence). *Suppose* (N, A) *is a strongly connected network. Suppose that for all* $p \in P$, $T_p : R_+^{n_1} \to R_+^1$ *is a continuous function and that, for all* $h \in R_+^{n_1}$, $T_p(h) > \varepsilon$ *for some* $\varepsilon \geqq 0$. *Also suppose that for all* $i \in I$, $D_i : R_+^{n_2} \to R_+^1$ *is a nonnegative continuous function that is bounded from above on the set* $\{u \in R^{n_2} : u_i \geqq \varepsilon \text{ for all } i\}$. *Then the traffic equilibrium system* (3.1) *has a solution.*

Proof. Let $\hat{e}$ be a vector of ones with $|I|$ components, and define

$$T_p'(h) = T_p(h) - \varepsilon > 0, \qquad D_i'(u) = D_i(u + \varepsilon \hat{e}).$$

These functions satisfy the hypothesis of Theorem 5.4, and so they are guaranteed to have a complementarity (or equilibrium) solution (h', u'). But then $(\hat{h}, \hat{u}) = (h', u' + \varepsilon \hat{e})$ is a complementarity (equilibrium) solution for T_p and D_i. □

6. Uniqueness. In situations in which the traffic equilibrium problem can be formulated as an equivalent convex optimization problem, the Kuhn–Tucker vector associated with the flow constraints $\sum_{p \in P_i} h_p = D_i(u)$ can be identified with the vector u of shortest travel times (generalized costs). Since the gradients of these constraints as i varies are linearly independent, the theory of convex optimization implies that in equilibrium the shortest travel times are unique even if the flow vector h is not unique. This situation reflects practice as well. Generally, flow patterns in urban transportation networks vary, sometimes considerably, from day to day though travel times remain essentially constant.

In this section, we show that these observations apply to the additive version (3.1) and (3.2) of the general traffic equilibrium model as well. We first recall conditions due to Asmuth [6] that insure that link flows and shortest travel times are both unique. We

then show that imposing weaker conditions will still imply that shortest travel times are unique.

To facilitate our discussion in this section, we represent the traffic equilibrium problem in a matrix form. Let v_a denote the total flow on arc a, that is, $v_a = \sum_{i \in I} \sum_{p \in P_i} \delta_{ap} \cdot h_p$, and let v with dimension $|A|$ denote the vector of arc flows. Since we are assuming an additive model, $T_p(h) = \Delta^T t(h) = \sum_{a \in A} \delta_{ap} t_a(h)$ for every path $p \in P$. In fact, we will assume that the arc delay term $t_a(h)$ can be expressed as a function of link flows v and write $t_a(v)$.

Also, let $t(v)$ be the vector of arc delay functions and $D(u)$ be the vector of demand functions. Recall that $\Delta = (\delta_{ap})$ is the arc-path incidence matrix with dimension $|A| \times n_1$. Let $\Gamma = (\gamma_{pi})$ be the path O-D pair incidence matrix with dimension $n_1 \times n_2$; i.e., $\gamma_{pi} = 1$ when path p joins O-D pair i and $\gamma_{pi} = 0$ otherwise.

Then the traffic-equilibrium problem can be written as

$$
\begin{aligned}
&(\Delta^T \cdot t(\Delta h) - \Gamma \cdot u) \cdot h = 0, \\
&\Delta^T \cdot t(\Delta h) - \Gamma \cdot u \geqq 0, \\
&\Gamma^T \cdot h - D(u) = 0, \\
&h \geqq 0, \qquad u \geqq 0.
\end{aligned} \tag{6.1}
$$

Now let $x = (h, u)^T$ and let $F(x): R_+^n \to R^m$ be defined as in § 4 as $F(x) = (\Delta^T t(\Delta h) - \Gamma u, \Gamma^T h - D(u))$. Then (4.1) is the nonlinear complementarity version (4.2) of (6.1).

Whenever $F(x)$ is strictly monotone, the solution to the general nonlinear complementarity problem (4.1) is unique (see Theorem 5.2). Asmuth [6] has extended this result to establish the following uniqueness result, which we state without proof.

THEOREM 6.1. (Uniqueness). *For a strongly connected network (N, A) suppose that t, the vector of the volume delay functions, and $-D$, the vector of the negative of the demand functions, are strictly monotone. Then the arc volumes v and the accessibility vector u for the additive traffic equilibrium problem* (3.1) *and* (3.2) *are unique, and the set of equilibrium path flows is convex.*

Observe the distinction between the hypothesis of this theorem and the assumption that $F(x)$ is strictly monotone. The theorem requires that the vector t of volume delay functions be strictly monotone in terms of arc volumes v whereas the latter assumption requires strict monotonicity in terms of path flows h. As we have noted earlier, the path flows need not be unique since two collections of path flows might correspond to the same arc flows.

Note that to insure the uniqueness of (v, u), Theorem 6.1 requires that both of the functions t and $-D$ are strictly monotone. Our next result shows that the strict monotonicity of $-D$ can be relaxed and, moreover, that uniqueness of u is maintained if either t or $-D$ is strictly monotone.

THEOREM 6.2. *For a complete network (N, A) suppose that t and $-D$ in the additive traffic equilibrium problems* (3.1) *and* (3.2) *are both monotone functions. If either of t or $-D$ is strictly monotone, then u is unique. Also, if t is strictly monotone and D is a positive function, then (v, u) is unique.*

Proof. Suppose that $x^1 = (h^1, u^1)$ and $x^2 = (h^2, u^2)$, $x^1 \neq x^2$, are two solutions to the equilibrium problem. Nonnegativity of x^1, x^2, $F(x^1)$, and $F(x^2)$ and the complementarity conditions $x^1 F(x^1) = 0$ and $x^2 F(x^2) = 0$ imply that

$$(x^1 - x^2)[F(x^1) - F(x^2)] \leqq 0$$

or, substituting $(h, u)^T$ for x and using the definition of F,

$$(h^1-h^2)^T(\Delta^T t(\Delta h^1)-\Gamma u^1-\Delta^T t(\Delta h^2)+\Gamma u^2)$$
$$+(u^1-u^2)^T(\Gamma^T h^1-D(u^1)-\Gamma^T h^2+D(u^2))\leqq 0$$

or,

$$(\Delta h^1-\Delta h^2)^T(t(\Delta h^1)-t(\Delta h^2))+(u^1-u^2)^T(-D(u^1)+D(u^2))\leqq 0. \tag{6.2}$$

But both t and $-D$ are monotone functions, and thus each term in (6.2) is zero; that is,

$$(\Delta h^1-\Delta h^2)^T(t(\Delta h^1)-t(\Delta h^2))=0 \tag{6.3}$$

and

$$-(u^1-u^2)^T(D(u^1)-D(u^2))=0. \tag{6.4}$$

If $-D$ is strictly monotone, then (6.4) implies that $u^1=u^2$, or u is unique.

Now, suppose that t is strictly monotone. Then (6.3) implies that $v^1=\Delta h^1=\Delta h^2=v^2$, or that the arc volume vector v is unique. But then the travel time, $t_a(v)$, on each arc is unique, which implies from (3.1a) and (3.2) that u is unique, since D being positive implies that some path flow is positive for each O-D pair i. □

Whenever t_a is a function only of the total volume in the arc, as when all the traffic from different origins has the same effect on the travel time of each arc, and there is no interaction between opposing lanes of two-way traffic or right or left turn penalties, then the strictly monotone condition on t can be relaxed for the uniqueness results.

COROLLARY 6.1. (Special case). *For a strongly connected network (N, A), suppose that each t_a of the additive traffic equilibrium problems* (3.1) *and* (3.2) *is a function only of v_a, and that it is monotone. Also, suppose that $-D$ is monotone and negative. Then u is unique.*

Proof. By definition t, the vector of the volume delay function, is monotone because each of its components is monotone. Thus (6.3) in the proof of Theorem 6.2 is valid. But since each component of t is monotone, (6.3) can be separated into a single term for each arc:

$$(v_a^1-v_a^2)(t_a(v_a^1)-t_a(v_a^2))=0 \quad \text{for all } a\in A.$$

This equation implies that $t_a(v_a^1)=t_a(v_a^2)$, or that the travel time on each arc is unique and, consequently, that u, the minimum path travel time, is unique. □

REFERENCES

[1] H. Z. AASHTIANI, *The Multi-modal Traffic Assignment Problem*, U.S. Dept. of Transportation, Tech. Summ. 77-1, January 1977.

[2] ———, *The Multi-modal Traffic Assignment Problem*, Ph.D. Dissertation, Operations Research Center, MIT, Cambridge, MA, May 1979.

[3] H. Z. AASHTIANI AND T. L. MAGNANTI, *Modeling and Computing Extended Urban Traffic Equilibria*, Final Report, U.S. Dept. of Transportation, draft, September 1979.

[4] M. ABDULAAL AND L. J. LEBLANC, *Methods for combining modal split and equilibrium assignment models*, Transport Sci., 13 (1979), pp. 292–314.

[5] B. AHN, *Computation of Market Equilibrium for Policy Analysis: The Project Independence Evaluation Systems Approach*. Ph.D. Thesis, Dept. of Engineering Economic Systems, Stanford University, Stanford, CA, May 1978.

[6] R. L. ASMUTH, *Traffic Network Equilibrium*. Tech. Rept. SOL-78-2, Stanford University, Stanford, CA, 1978.

[7] M. J. BECKMAN, C. B. MCGUIRE AND C. B. WINSTEN, *Studies in the Economics of Transportation*, Yale University Press, New Haven, CT, 1956.

[8] D. P. BERTSEKAS, *Algorithms for Optimal Routing of Flow in Networks*, Coordinated Sciences Laboratory Working Paper, Univ. of Illinois at Champaign-Urbana, 1976.
[9] D. P. BERTSEKAS AND E. M. GAFNI, *Projection Methods for Variational Inequalities with Applications to the Traffic Assignment Problem*, Laboratory for Information and Decision Systems, MIT, Cambridge, MA, Spring 1980.
[10] D. BRAESS AND G. KOCH, *On the existence of equilibria in asymmetrical multiclass-user transportation networks*, Transport Sci., 13 (1979), pp. 56–63.
[11] M. BRUYNOOGHE, A. GIBERT AND M. SAKAROVITCH, *Une méthode d'affectation du trafic*, 4th Symposium on Theory of Traffic Flow, Karlsruhe, 1968.
[12] *Chicago Area Transportation Study*, Final Rept., vol. 2, 1960, pp. 104–110.
[13] S. C. DAFERMOS, *An extended traffic assignment model with applications to two-way traffic*, Transport. Sci., 5 (1971), pp. 366–389.
[14] ———, *The General Multimodal Network Equilibrium Problem with Elastic Demand*, Lefschetz Center for Dynamical Systems, Brown University, Providence, RI, Spring 1980.
[15] ———, *The traffic assignment problem for multiclass-user transportation networks*, Transport Sci., 7 (1972), pp. 73–87.
[16] ———, *Traffic equilibrium and variational inequalities*, Transport. Sci., 14 (1980), pp. 42–54.
[17] S. C. DAFERMOS AND F. T. SPARROW, *The traffic assignment problem for a general network*, J. Research NBS Ser. B., 73B (1969), pp. 91–118.
[18] R. S. DEMBO AND J. G. KLINCEWICZ, *An Approximate Second-Order Algorithm for Network Flow Problems with Convex Separable Costs*, Yale School of Organization and Management, Working Paper Series No. 21, 1978.
[19] *Detroit Area Transportation Study*, Vol. 2, 1958, pp. 79–107.
[20] R. B. DIAL, *A combined trip distribution and modal split model*, Paper presented at the 1974 Annual Meeting, Highway Research Board, 1973.
[21] T. A. DOMENEICH AND D. MCFADDEN, *Urban Travel Demand*, North-Holland, Amsterdam, 1975.
[22] M. FLORIAN, *A traffic equilibrium model of travel by car and public transit modes*, Transport. Sci., 11 (1977), pp. 166–179.
[23] M. FLORIAN AND S. NGUYEN. *A method for computing network equilibrium with elastic demand*, Transport. Sci., 8 (1974), pp. 321–332.
[24] ———, *Recent experience with equilibrium methods for the study of a congested urban area*, Proc. International Symposium on Traffic Equilibrium Methods, University of Montreal, 1974.
[25] ———, *An application and validation of equilibrium trip assignment methods*, Transport. Sci., 10 (1976), pp. 376–390.
[26] B. GOLDEN, *A minimum-cost multi-commodity network flow problem concerning imports and exports*, Networks, 5 (1975), pp. 331–356.
[27] D. W. HEARN AND H. KUHN, *Network Aggregation in Transportation Planning—Final Report*, Mathtech, Inc., Princeton, NJ, 1977.
[28] N. A. IRWIN, N. DODD AND H. G. VON CUBE, *Capacity restraint in assignment programs*, Highway Res. Board, 297 (1961), pp. 109–127.
[29] N. A. IRWIN AND H. G. VON CUBE, *Capacity restraint in multi-travel mode assignment programs*, Highway Res. Board, 347 (1962), pp. 258–289.
[30] J. JACOBSON, *Case Study Comparison of Alternative Urban Travel Forecasting Methodologies*, S. M. Dissertation, Dept. of Civil Engineering, MIT, Cambridge, MA, 1977.
[31] S. KARAMARDIAN, *The nonlinear complementarity problem with applications, parts* I *and* II. J. Optim. Theor. Appl., 4 (1969), pp. 87–98.
[32] ———, *The complementarity problem*, Math. Programming, 2 (1972), pp. 107–129.
[33] M. KOJIMA, *A unification of the existence theorems of the nonlinear complementarity problem*, Math. Programming, 9 (1975), pp. 257–277.
[34] L. J. LEBLANC, *Mathematical Programming Algorithms for Large Scale Network Equilibrium and Network Design Problems*. Ph.D. Dissertation, Dept. Industrial Engineering and Management Science, Northwestern Univ., Evanston, IL, 1973.
[35] L. J. LEBLANC, E. MORLOK AND W. PIERSKALLA, *An efficient approach to solving the road network equilibrium traffic assignment problem*, Transport Sci., 9 (1975), pp. 309–318.
[36] T. L. LEVENTHAL, G. L. NEMHAUSER AND L. E. TROTTER, *A column generation algorithm for optimal traffic assignment*, Transport. Sci., 7 (1973), pp. 168–176.
[37] M. L. MANHEIM AND E. R. RUITER, DODOTRANS I: *A decision oriented computer language for analysis of multi-mode transportation systems*, Highway Res. Record, 314 (1970), pp. 135–163.
[38] B. V. MARTIN AND M. L. MANHEIM, *A research program for comparison of traffic assignment techniques*, Highway Res. Record, 88 (1965), pp. 69–84.

[39] B. V. MARTIN, F. W. MEMMOTT AND A. J. BONE, *Principles and Techniques of Predicting Future Demand for Urban Area Transportation*, Res. Rpt. R63-1, Civil Engineering Department, MIT, Cambridge, MA, 1963.

[40] W. L. MERTZ, *Review and evaluation of electronic computer traffic assignment programs*, Highway Res. Board, 297 (1961), pp. 94–105.

[41] S. NGUYEN, *Une Approche Unifiée des Méthodes d'Équilibre pour l'Affectation du Trafic*, Ph.D. Dissertation, Pub. 171, Dept. d'Informatique, Université de Montréal, 1973.

[42] ———, *An algorithm for the traffic assignment problem*, Transport. Sci., 8 (1974), pp. 203–216.

[43] ———, *A Mathematical Programming Approach to Equilibrium Methods of Traffic Assignment with Fixed Demands*. Pub. 17, Centre de Recherche sur les Transports, Université de Montréal, 1976.

[44] ———, *Equilibrium Traffic Procedures with Elastic Demands*, Pub. 39, Centre de Recherche sur les Transports, Université de Montréal, 1976.

[45] S. NGUYEN AND L. JAMES, TRAFFIC—*An Equilibrium Traffic Assignment Program*, Pub. 17, Centre de Recherche sur les Transports, Université de Montréal, 1975.

[46] Y. SHEFFI, *Transporation Networks Equilibrium with Discrete Choice Models*, Ph.D. Dissertation, Dept. of Civil Engineering, MIT, Cambridge, MA, 1978.

[47] M. L. SMITH, *The existence, uniqueness and stability of traffic equilibria*, Transport. Res., 13B (1979), pp. 295–304.

[48] R. SMOCK, *A comparative description of a capacity-restrained traffic assignment*, Highway Res. Record, 6 (1963), pp. 12–40.

[49] P. A. STEENBRINK, *Optimization of Transport Networks*, Wiley, London, 1974.

[50] M. J. TODD, *The Computation of Fixed Points and Applications*. Lecture Notes in Economics and Mathematical Systems, 124, Springer-Verlag, Berlin, 1976.

[51] UMTA *Transportation Planning System* (UTPS) *User's Guide*, U.S. Dept. of Transportation, Urban Mass Transit Administration, Office of Transit Planning, Planning Methodology and Technical Support Division, Washington, DC, 1976.

[52] J. G. WARDROP, *Some theoretical aspects of road traffic research*, Proc. Inst. Civil Engineers, Part II, 1 (1952), pp. 325–378.

[53] D. K. WITHEFORD, *Traffic assignment analysis and evaluation*, Highway Res. Record, 6 (1963), pp. 1–11.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 3, September 1981

0196-5212/81/0203-0002 $01.00/0

GRAPH THEORETIC METHODS FOR THE QUALITATIVE ANALYSIS OF RECTANGULAR MATRICES*

HARVEY J. GREENBERG†, J. RICHARD LUNDGREN‡ AND JOHN S. MAYBEE§

Abstract. In the past few years, many large models including several energy models have been represented by rectangular matrices, and graphs appear to be valuable in investigating connectivity and other properties of these models. It is the purpose of this paper to establish some of the basic foundations for the use of graphs and digraphs to investigate properties of rectangular matrices. A variety of graphs and digraphs associated with rectangular matrices are introduced, and several theorems related to connectivity and tearing are proved. There are also a few applications to the area of computer-assisted analysis.

Introduction. In the past 25 years considerable use has been made of the relationships between graphs and square matrices. The theory of the use of square matrices to analyze digraphs is developed in Harary, Norman and Cartwright [14]. More recently graphs and digraphs have been used in research on square matrices in sparse matrix theory (see [3], [20], [21], [22], [23], [24]) and qualitative matrix theory (see Maybee and Quirk [16]). It is the purpose of this paper to establish some of the basic foundations for the use of graphs and digraphs to investigate properties of rectangular matrices. Although some use has been made of some of these graphs in the past (see Dulmage and Mendelsohn [4], Tewarson [17] and Weil and Kettler [19]), it is the development of computer-assisted analysis (CAA) for matricial forms (see Greenberg [6], [7], [8]) that has created a need for a comprehensive study of these graphs.

In these papers, Greenberg develops the basic concepts associated with CAA and matricial forms and describes their use in model simplification. He demonstrates that graphs are valuable in investigating connectivity and other properties of matricial forms, which for our purposes will be treated simply as rectangular matrices. For some models, we know each matrix entry, for others, only the sign of each entry, and for others, only the locations of the nonzeros. So, we are using some graphs where only the locations of the nonzeros is needed and others where the sign of the entries is needed.

Our general approach is to develop the theory of graphs associated with rectangular matrices. However, we have included some applications to CAA. For further applications, see Greenberg, Lundgren and Maybee [9], where we presented in an expository paper several applications of this theory to CAA, and [10], where applications are presented in an operations research context. In the first section we define the basic graphs and digraphs associated with rectangular matrices. Then in the next two sections we present several theorems related to connectivity and tearing.

1. Graphs of rectangular matrices. Given a rectangular matrix M, we define two sets of vertices, $R = \{r_1, \cdots, r_m\}$ and $C = \{c_1, \cdots, c_n\}$, to represent the row and column variables, respectively. The three basic undirected graphs are:

Fundamental bigraph. B is a bipartite graph on R, C. The edges E correspond to the nonzeros in M: $[r_i, c_j]$ is in E if and only if $M(i, j) \neq 0$.

Row graph. RG is defined on R. Its edges are defined by: r_i and r_k are adjacent if there exists c_j in C such that $[r_i, c_j]$ and $[r_k, c_j]$ are in E. In other words, two rows are adjacent if they have a common column intersection in M.

* Received by the editors September 9, 1980, and in revised form January 15, 1981.

† Energy Information Administration, Washington, DC 20461.

‡ Allegheny College, Meadville, Pennsylvania 16335. The research of this author was supported by the National Science Foundation under grant SPI-7916608 while he was visiting the University of Colorado.

§ University of Colorado, Boulder, Colorado 80309.

Column graph. CG is defined on C. Its edges are defined by: c_j and c_k are adjacent if there exists r_i in R such that $[c_j, r_i]$ and $[c_k, r_i]$ are in E. In other words, two columns are adjacent if they have a common row intersection in M.

The row and column graphs are the "2-step" graphs recently studied by Exoo and Harary [5]. Their extension to "n-step" graphs, induced by the fundamental graph by paths of length n, may also prove valuable. For some applications it is useful to sign B, and if possible to sign RG or CG. This possibility is investigated in a related paper (see [11]).

To capture the information contained in the signs of the nonzeros, three basic digraphs are defined:

Fundamental digraph. D has the same points as B and the orientation of the arcs, A, is defined by the signs of the nonzeros:

$$[r_i, c_j] \text{ in } E \text{ and } M(i, j) < 0 \text{ iff } (r_i, c_j) \text{ in } A,$$

$$[r_i, c_j] \text{ in } E \text{ and } M(i, j) > 0 \text{ iff } (c_j, r_i) \text{ in } A.$$

Note that B is the undirected graph formed from D by deleting the directions of the arcs.

Row digraph. RD is defined on R. Its arcs are defined by: (r_i, r_k) is an arc if and only if there exists c_j in C such that (r_i, c_j) and (c_j, r_k) are in A. (The arcs of RD may not be isomorphic to the edges of RG.)

Column digraph. CD is defined on C. Its arcs are defined by: (c_j, c_k) is an arc if and only if there exists r_i in R such that (c_j, r_i) and (r_i, c_k) are in A. (The arcs of CD may not be isomorphic to the edges of CG.)

If we want to refer to the matrix M that a graph or digraph is associated with, we will use the notation $B(M)$.

The fundamental digraph is the signal flow graph, familiar in engineering science (see Henley and Williams [15]). It represents a flow concept, such as the "physical flows" matricial form (see Greenberg [6]). These graphs and digraphs are also useful in working with sparse matrices (see Duff [3]).

One additional concept is relevant to our study, namely "combivalence". This grew out of linear programming and was formalized as an algebra by Tucker [18]. Two matrices are *combivalent* if one is reachable from the other by a sequence of pivot operations. This relation is denoted M' comb. M, and associated graphs are also indicated by primes. It is not difficult to show that combivalence is reflexive, symmetric and transitive (see [18] where the proofs are given). We note that the pivot operation used in linear programming is similar to the total pivoting strategy used in Gauss elimination, except that the choice of a pivot is based upon somewhat different criteria. However, once a pivot is chosen, say a_{ij}, elements of the jth column are converted to zero except for a_{ij}, which is changed to 1, by exactly the same operations as those used in Gaussian elimination. For some examples see [6].

Our interest in combivalence stems from the changing topology of the basic graphs when reconfiguring the matrix form—that is, redefining which variables are in the row set and which are in the column set.

2. General connectivity. In this section we investigate connectivity relations among the basic graphs and digraphs. An understanding of connectivity helps to provide computer-assisted analysis (see Greenberg [6], [7], [8]). One associated CAA function is model verification.

Throughout this section we assume that M is an $m \times n$ matrix such that each column and row of M has a nonzero element.

The following lemma will be used throughout the paper.

LEMMA 2.1. *Suppose $M_1 = PMQ$, where P and Q are permutation matrices. Then each of the graphs for M_1 is isomorphic to the corresponding graph for M.*

The above result is clear since by permuting the rows and columns of M we relabel the graph but do not change its structure. Therefore, when we rearrange the rows and columns of M we will still refer to the matrix as M for simplicity of notation.

The next theorem establishes an important relationship between the three graphs; that is, if one is connected, then they all are. Combined with its corollary this tells us that each of the basic graphs have the same reachability.

THEOREM 2.2. *Let M be an $m \times n$ matrix such that each column and row of M has a nonzero element. The following are equivalent*:

1) *CG is connected.*
2) *RG is connected.*
3) *B is connected.*

Proof. First we will prove the equivalence of 1) and 2). Suppose CG is disconnected. Then there are column variables k_1 and $\tilde{k}_1$ that are not connected in CG. Suppose there are p column variables $k_1, \cdots, k_p$ connected to k_1, then interchange the columns of M so that $k_1, \cdots, k_p$ form the first p columns of M. Hence, $\tilde{k}_1 \neq k_i$, $i = 1, \cdots, p$. Now, for each row, $i = 1, \cdots, m$, if $m_{il}, \cdots, m_{ip}$ are all zero, interchange it with the row closest to the bottom that has a term $m_{ij} \neq 0$ for some j, $j = 1, \cdots, p$. Also interchange columns so that $\tilde{k}_1$ is the $p+1$ column of M. We now have the following block form for M:

$$\begin{array}{c} \begin{matrix} k_1 \cdots k_p & \tilde{k}_1 \end{matrix} \\ \left[\begin{array}{c|c} M_{11} & M_{12} \\ \hline 0 & M_{22} \end{array}\right], \end{array}$$

where 0 is a zero-matrix and each row of M_{11} has a nonzero entry. Now suppose some entry of M_{12} is nonzero, say in row q. Then row q has a nonzero entry in some column k_j of M_{11}, and also in some column c_r of M_{12}. But then c_r is adjacent to k_j, so c_r is connected to k_1, a contradiction. Our matrix now has the form

$$\begin{array}{c} \begin{matrix} k_1 \cdots k_p & \tilde{k}_1 \end{matrix} \\ \left[\begin{array}{c|c} M_{11} & 0 \\ \hline 0 & M_{22} \end{array}\right], \end{array}$$

where M_{11} is $s \times p$ and M_{22} is $(m-s) \times (n-p)$. Since column $\tilde{k}_1$ has a nonzero entry, it must be in M_{22}, say in row t of M. Then $t > s$ and row t is not connected to any row of M_{11}, so RG is disconnected.

Now suppose RG is disconnected. Then there are now variables r_1 and $\tilde{r}_1$ that are not connected, and then we can rearrange the columns and rows of M so it has the form

$$\begin{array}{c} \begin{matrix} \\ r_1 \\ \vdots \\ \tilde{r}_s \\ \\ \tilde{r}_1 \end{matrix} \\ \left[\begin{array}{c|c} & \\ M_{11} & 0 \\ & \\ \hline & \\ M_{22} & M_{22} \\ & \end{array}\right], \end{array}$$

where $\{r_1, \cdots, r_s\}$ are all the rows connected to r_1 and each column of M_{11} has a nonzero entry. Then we see that each row of M_{21} must have only zeros, so that our rearranged

matrix has the form

$$\begin{matrix} r_1 \\ \vdots \\ r_s \\ \\ \tilde{r}_1 \end{matrix} \begin{bmatrix} & \\ M_{11} & 0 \\ & \\ \hline & \\ 0 & M_{22} \\ & \end{bmatrix},$$

where M_{11} is $s \times p$ and M_{22} is $(m-s)\times(n-p)$. Since $\tilde{r}_1$ has a nonzero entry in M_{22}, we see that for any column c_t with $t > p$, c_t is not connected to any columns of M_{11}, so CG is disconnected.

We complete the proof by showing the equivalence of 3) and 1). Suppose CG is disconnected. Then it is easy to see that B is disconnected by examining the rearranged matrix above. Now suppose B is disconnected. Then there are points x_1, x_2 such that there is no path between x_1 and x_2. There are three possibilities to consider: they are both row points, both column points, or one is a row point and one is a column point.

Suppose row points r_1 and r_2 are not connected in B. If r_1 and r_2 are connected in RG, then there exists a path and hence a shortest path $\{r_{i1} = r_1, r_{i2}, \cdots, r_{ip} = r_2\}$ from r_1 to r_2 in RG. Then since r_{i1} is adjacent to r_{i2}, there exists a column c_{j1} such that c_{j1} has a nonzero entry in both r_{i1} and r_{i2}. Similarly, r_{i_k} adjacent to $r_{i_{k+1}}$ implies there is a column c_{j_k} such that c_{j_k} has nonzero entries in r_{i_k} and $r_{i_{k+1}}$. Hence, we get a sequence of rows and columns $\{r_{i_1}, c_{i_1}, r_{i_2}, c_{i_2}, \cdots, c_{i_{p-1}}, r_{i_p}\}$ which determines a path in B between r_1 and r_2, a contradiction. Hence, RG is disconnected and so CG is disconnected.

Similarly, if column points c_1 and c_2 are not connected in B we get that CG is disconnected.

Finally, to complete the proof, we only have to consider the case where all rows of M are connected in B, all columns are connected in B, but some row r and column c are not connected. If there is only one column, then since each row has a nonzero entry, we would have B connected, so we may assume there is more than one column. Since c is connected to every column, then c and some other column c_1 must intersect a row r_1, so c is adjacent to r_1. But r_1 is connected to r, so c is connected to r, a contradiction.

Hence, if B is disconnected then CG is disconnected. The proof is complete.

Theorem 2.2 has the following important corollary.

COROLLARY 2.3. *Let M be an $m \times n$ matrix such that each column and row of M has a nonzero element. The following hold*:

1) *Each of the graphs B, RG and CG has the same number of components.*
2) *The rows and columns of M can be rearranged so that M is in block diagonal form with each diagonal block corresponding to a component.*

Proof. The proof is by induction on the number of components N. If $N = 1$, the result holds by Theorem 2.2. So assume the result holds for all matrices M where CG has less than N components.

Suppose CG has N components, $N > 1$. Then there are column variables k_1 and $\tilde{k}_1$ that are not connected, so we can rearrange the rows and columns of M as in the proof of Theorem 2.2 so that M has the following form:

$$\begin{bmatrix} E & 0 \\ 0 & F \end{bmatrix}.$$

In the above matrix E corresponds to the component of CG containing k_1, and since $CG(E)$ is connected, so is $B(E)$ and $RG(E)$. Since $CG(M)$ has N components,

$CG(F)$ has $N-1$ components, so by induction $B(F)$ and $RG(F)$ have $N-1$ components. Since the number of components of $B(M)$ and $RG(M)$ is just the sum of the components of E and F, we have that 2) holds. Finally, since F has $N-1$ components, by induction it can be put in block diagonal form with each block corresponding to a component, and so we then have M in block diagonal form.

By the above theorem and corollary it is reasonable to say that M has N components if $CG(M)$ has N components and that M is connected if $N=1$. Now we turn our attention to combivalence. It turns out that the component structure is unaffected by pivoting, and two columns (rows) in the same component of M must be adjacent in some M' combivalent to M.

THEOREM 2.4. *Let M be an $m \times n$ matrix such that each column and row of M has a nonzero element. Given i and j in $C(R)$, the following are equivalent:*

1) *i and j are in the same component of B.*
2) *i and j are in the same component of $CG(RG)$.*
3) *i and j are in the same component of B' for all M'* comb. *M.*
4) *There exists M'* comb. *M for which i and j are adjacent in $CG'(RG')$.*

Proof. First note that 1) and 2) are equivalent by the previous theorem and corollary. Also, since 3) implies 1), we have that 3) implies 2). Now suppose i and j are in the same component of CG. If i and j are in different components of B' for some M' comb. M, then they are also in different components of CG'. By the argument used in Theorem 2.2 we can rearrange M' into a block diagonal form

$$\left[\begin{array}{c|c} A' & 0 \\ \hline 0 & B' \end{array}\right],$$

which separates i and j by having i as a column of A' and j as a column of B'. Every pivot retains the block diagonal form, so no configuration can connect the two variables. However, since combivalence is reflexive, M comb. M', and i and j are connected in M, a contradiction. A similar argument works if we assume i and j are in the same component of RG.

Next we establish the equivalence of 2) and 4), again working with CG. First note that if 4) holds, then i and j are in the same component of CG', and since M comb. M', i and j are in the same component of CG. Hence, 4) implies 2).

Suppose i and j are in the same component of CG; then i and j are connected by a path, and hence a shortest path $i = u_1, i_2, \cdots, i_{r+1} = j$ of length r. We can then rearrange the columns of M so that these $r+1$ columns are the first $r+1$ columns of M, and so without loss of generality we may assume that $i=1$ and $j=r+1$. Furthermore, we can rearrange the rows of M so that the adjacencies occur in the first r rows of M. We now have a shortest path submatrix for the path from 1 to $r+1$ in the upper left-hand corner of M as illustrated in Fig. 1.

Observe that the $r \times (r+1)$ shortest path submatrix has nonzeros on the main diagonal and superdiagonal and zeros elsewhere.

Now we want to show that 1 is adjacent to $r+1$ for some matrix M' comb. M. If $r=1$ we are done, since 1 is adjacent to $r+1$ in M. So suppose $r>1$. Let A be the $r \times (r+1)$ matrix in the upper left-hand corner of M representing the shortest path. We will just consider the effects of pivoting on A. Since $r>1$, $a_{22} \neq 0$ and $a_{23} \neq 0$, so we will pivot on a_{22}, thus changing the configuration by pivoting the second column into the row set. If we let $B = A'$, then $b_{11} = a_{11} - (a_{12}/a_{22}) \cdot 0 = a_{11} \neq 0$ and $b_{13} = a_{13} - (a_{12}/a_{22}) \cdot a_{23} = (-a_{12}/a_{22}) \cdot a_{23} \neq 0$. Hence in B we have that 1 is adjacent to 3, and the nonzeros in the other rows and columns of B remain as in A, so in B we have a path from 1 to $r+1$ of length $r-1$. Repeated application of this pivoting strategy results

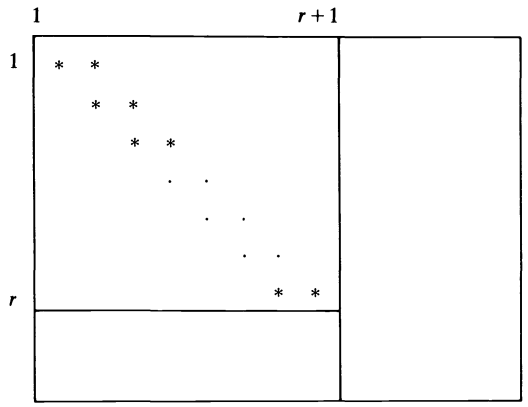


FIG. 1. *Shortest path connecting two column variables.*

in a path of length 1. Hence, we can find M' comb. M for which 1 is adjacent to $r+1$ in CG', so we have shown that 2) implies 4).

The following result follows immediately from Theorem 2.4.

COROLLARY 2.5. *Let M be an $m \times n$ matrix such that each column and row of M has a nonzero element. Then B has N components if and only if B' has N components for all M'* comb. *M.*

While eventual reachability, in the sense of component structure, is the same in all three basic graphs, and is invariant under pivoting, other connectivity properties are not. In particular, the density of only one row (column) may dilute certain structures contained in the fundamental bigraph when we consider the column (row) graph. The following theorem, for example, shows that the presence of one dense row, such as a constraint on the aggregate activity level which might occur in certain models represented by our matrix, renders the column graph complete; block structures, associated with regions in the model, become hidden when we examine the column graph.

THEOREM 2.6. *If a column (row) has k nonzeros, then $RG(CG)$ has a complete subgraph with k vertices.*

Proof. Suppose column j has k nonzeros in rows $i_1, \cdots, i_k$. Then in RG, the rows $r_{i1}, \cdots, r_{ik}$ are adjacent and hence RG has a complete subgraph with k vertices. The proof is similar if a row has k nonzeros.

One criterion that helps to decide on the choice of graphs to use in investigating such questions as the possible block diagonal form of A (see Corollary 2.3), is the complexity of the graphs. A measure of this is the number of edges in each of the graphs. Even if we cannot actually compute these numbers, it may prove useful to have some estimates for them. Our results are far from complete, but they go in the proper direction because they estimate these numbers for CG and RG in terms of the number of edges in B, a number which is easy to compute. We hope that a more thorough investigation of this problem will produce sharper results, especially for the case where B has 4-cycles.

Next we find relationships between $|E_C|$, $|E_R|$ and $|E_B|$, the number of edges in CR, RG and B respectively. First note that $|E_B| = z$, the number of nonzeros in M. We make the following definitions:

$c(j)$ = number of nonzeros in column j,

$r(i)$ = number of nonzeros in row i,

$M_C = \max_{1 \leqq j \leqq n} c(j)$,
$M_R = \max_{1 \leqq i \leqq m} r(i)$.

THEOREM 2.7. *If B has no 4-cycles, then*

$$|E_C| = \sum_{r(i) \geqq 2} \frac{r(i)(r(i)-1)}{2} \quad \text{and} \quad |E_R| = \sum_{c(j) \geqq 2} \frac{c(j)(c(j)-1)}{2}.$$

Proof. First we determine $|E_C|$. Suppose columns c_i and c_j have nonzeros in rows r_p and r_q. Then r_p, c_i, r_q, c_j, r_p is a 4-cycle in B, a contradiction. Hence, each pair of columns has common nonzeros in at most one row. From this it follows that each pair of nonzeros in any row determines a unique edge.

If $r(i) \leqq 1$, then there are no edges in CG determined by row r_i.

If $r(i) \geqq 2$, the number of pairs of nonzeros in row r_i and hence the number of edges in CG determined by row r_i is

$$\frac{r(i)(r(i)-1)}{2}.$$

We get $|E_C|$ by summing over those rows with more than one nonzero.

The formula for $|E_R|$ is derived in a similar way.

One implication of Theorem 2.7 is that we can get lower bounds for $|E_C|$ and $|E_R|$ in terms of $|E_B|$.

COROLLARY 2.8. 1) *Suppose B has no 4-cycles, $r(i) \geqq 2$ for every row, and z_r = average row degree in B. Then*

$$|E_C| \geqq \tfrac{1}{2}|E_B|(z_r - 1) \geqq \tfrac{1}{2}|E_B|.$$

2) *Suppose B has no 4-cycles, $c(j) \geqq 2$ for every column, and z_c = average column degree in B. Then*

$$|E_R| \geqq \tfrac{1}{2}|E_B|(z_c - 1) \geqq \tfrac{1}{2}|E_B|.$$

Proof. 1) Observe that since $r(i) \geqq 2$ for every row, by Theorem 2.7 we have

$$|E_C| = \sum_{i=1}^{m} \frac{r(i)(r(i)-1)}{2},$$

where $\sum_{i=1}^{m} r(i) = z$. We get a lower bound for $|E_C|$ by solving the following problem:

$$\min \frac{1}{2} \sum_{i=1}^{m} x_i(x_i - 1) \quad \text{subject to} \sum_{i=1}^{m} x_i = z.$$

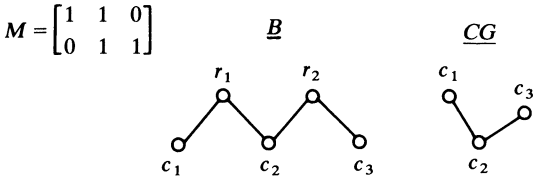
The solution is $x_i^* = z/m$ for all i. Hence, we have

$$\begin{aligned} |E_C| = \sum_{i=1}^{m} \frac{r(i)(r(i)-1)}{2} &\geqq \frac{1}{2} \sum_{z=1}^{m} \frac{z}{m}\left(\frac{z}{m} - 1\right) \\ &= \frac{1}{2} \cdot m \cdot \frac{z}{m}(z_r - 1) \\ &= \frac{1}{2}|E_B|(z_r - 1). \end{aligned}$$

since $z_r \geqq 2$ by our assumption, we have

$$|E_C| \geqq \frac{1}{2}|E_B|.$$

2) The proof for 2) is similar to the proof for 1). The lower bound can be attained, as illustrated for $|E_C|$ in Fig. 2.



$|E_C| = 2 = \frac{1}{2}|E_B|$.

FIG. 2

An implication of Corollary 2.8 is that, if the number of edges dominate the storage requirements, the fundamental bigraph uses less space than the column graph. To see this note that $|E_c| > |E_B|$ when the average row degree, Z_r, is greater than 3, which is generally the case. The row graph, however, may be sparser because many problems have an average column degree less than 3. One class of examples is the network problems, wherein every column has two nonzeros.

The next theorem provides general upper bounds on the number of edges in *RG* and *CG*, respectively.

THEOREM 2.9.

1) $$|E_C| \leqq \frac{(M_R - 1)}{2} |E_B|.$$

2) $$|E_R| \leqq \frac{(M_C - 1)}{2} |E_B|.$$

Proof.

1) $$|E_C| \leqq \sum_{i=1}^{n} \frac{r(i)(r(i)-1)}{2} \leqq \sum_{i=1}^{n} r(i) \frac{(M_R - 1)}{2}$$
$$= \frac{(M_R - 1)}{2} |E_B|.$$

The proof for 2) is similar.

Again Fig. 2 illustrates that the upper bounds can be attained. In fact, if B has no 4-cycles and every row has the same degree—that is, $r(i) = z/n$ for all i—then $z_r = M_R$; so we have

$$\tfrac{1}{2}|E_B|(z_r - 1) \leqq |E_C| \leqq \tfrac{1}{2}|E_B|(z_r - 1),$$

so equality holds throughout. Moreover, $z_c = M_C = 2$, as when M is an incidence matrix, implies

$$|E_R| = \tfrac{1}{2}|E_B|.$$

Now let us consider the basic digraphs.

THEOREM 2.10. *The following are equivalent*:

1) *D is strongly connected.*

2) *RD and CD are strongly connected.*

Proof. That 1) implies 2) follows from the definitions of D, RD, and CD. So suppose 2) holds. If we pick two rows r_i and r_j, then we can find a path from r_i to r_j and one from r_j to r_i in D by using the definition of RD and that RD is strongly connected. Similarly for two columns c_i and c_j we can find a path from c_i to c_j and from c_j to c_i since CD is strongly connected.

Now we need to show that given a column c and a row r there exists a path from c to r and a path from r to c in D. First observe that there must be at least 2 columns for RD to be strong and at least 2 rows for CD to be strong. Since there is a path from c to every other column in CD, there must be a column c_j that c is adjacent to in CD, hence there is a row r_i such that the arcs (c, r_i) and (r_i, c_j) are in D. Since there is a path from r_i to r in D and c is adjacent to r_i, we now have a path from c to r. A similar argument shows that there is a path from r to c. Hence D is strongly connected.

This provides directed reachability information analogous to Theorem 2.2, except that there is no equivalence with combivalent matrices. This is because, in general, signs change. Qualitative determinacy addresses the basic issues (see Greenberg [6]). However the following corollary to Theorem 2.10 is an analogy to part of Theorem 2.4.

COROLLARY 2.11. *The following are equivalent*:

1) *i and j are in the same strong component of $CD(RD)$.*

2) *i and j are in $C(R)$ and in the same strong component of D.*

Proof. The same ideas used in Theorem 2.10 establish the equivalence of 1) and 2).

We conclude this section by observing that the graph or digraph which is most useful depends on the application. For example, in [9] we show that in using a theorem of Bondy [2] to estimate the number of components of M, one generally gets a much better estimate using RG or CG than in using B. In [12] we show that B can be used to determine the singularity of certain square matrices and either RG or CG can be used to find the block diagonal form of M.

3. Tearing. The notion of tearing is to separate a model such that one portion satisfies a specified structure. One structure of interest is the block diagonal—that is, distinguish a set of vertices whose removal disconnects the graph. Such a set of vertices V is called an *articulation set.* A *cutset* of a graph is a set of edges whose removal disconnects the graph. In investigating the relationships among articulation sets in the basic graphs we find that articulation sets in B determine cutsets in RG or CG, and cutsets in RG or CG determine articulation sets in B.

The following lemma will be useful for proving the theorems that follow.

LEMMA 3.1. *Two column (row) vertices are connected by a path in B if, and only if, they are connected in $CG(RG)$ by a path containing the same column (row) vertices.*

Proof. We shall prove the lemma for column vertices. Suppose columns c_{j1} and c_p are connected by the path $c_{j1}, r_{i1}, c_{j2}, r_{i2}, \cdots, r_{ik}, c_p$ in B.

Since there is a path c_{j1}, r_{i1}, c_{j2} in B, then c_{j1} and c_{j2} are adjacent in CG. Similarly, c_{j2} and c_{j3} are adjacent in CG, and continuing this process, we get the path $\{c_{j1}, c_{j2}, \cdots, c_{jk}, c_p\}$ from c_{j1} to c_p in CG.

Clearly this process can be reversed if we start with a path in CG.

THEOREM 3.2. *Any articulation set for RG or CG is an articulation set for B.*

Proof. We shall prove the theorem for the case that R_0 is an articulation set for RG. Then there are vertices r_i and r_j that are not connected by any path in $RG - R_0$. Suppose that r_i and r_j are connected by a path in $B - R_0$. Then, by Lemma 3.1, they must be connected by a path in RG containing the same row vertices, and hence a path in $RG - R_0$, a contradiction. Hence, R_0 must contain an articulation set for B.

Let V be an articulation set for B. Are $V \cap R$ and $V \cup C$ articulation sets for RG and CG, respectively? Fig. 3 illustrates that the answer to this question may be no!

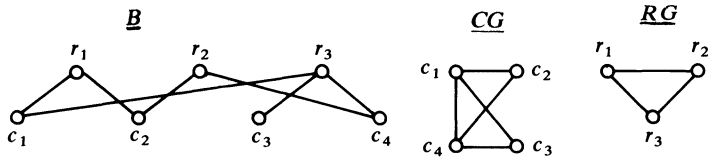


FIG. 3

$V = \{r_3\}$ is an articulation set for B, but $V \cap R = \{r_3\}$ is not an articulation set for RG.

When an articulation set is a singleton, its member is an *articulation vertex*. Let $\mathscr{A}_G$ be the number of articulation vertices for a graph G. By Theorem 3.2 we see that any articulation vertex for RG or CG is an articulation vertex for B, but Fig. 3 illustrates that an articulation vertex for B may not be an articulation vertex for either RG or CG. So we have the following result.

COROLLARY 3.3. $\mathscr{A}_B \geqq \mathscr{A}_{RG} + \mathscr{A}_{CG}$.

The next few theorems, which relate cutsets in the column or row graphs to articulation sets in the fundamental bigraph, may exploit Beineke and Harary's [1] integrated approach to separation. First, define $C(i, k)$ as the set of columns that intersect rows i and k. This is nonempty if and only if $[r_i, r_k]$ is an edge in RG. Similarly, define $R(j, k)$ for c_j, c_k in C. If F is any set of edges in RG or CG, then let $C(F)$ and $R(F)$ denote the unions of $C(i, k)$ or $R(j, k)$, respectively, for $[r_i, r_k]$ or $[c_j, c_k]$ in F.

THEOREM 3.4. *If F is a cutset for $RG(CG)$, then $C(F)(R(F))$ is an articulation set for B.*

Proof. Suppose F is a cutset for CG. Then there are column points c_p and c_q such that every path from c_p to c_q contains an edge from F. Suppose $[c_p, r_{i1}, c_{j2}, r_{i2}, \cdots, r_{it}, c_q]$ is a path from c_p to c_q in B. Then, by Lemma 3.1, $[c_p, c_{j2}, \cdots, c_{jt}, c_q]$ is a path from c_p to c_q in CG. Hence, an edge in this path, $[c_{jk}, c_{j(k+1)}]$, $\in F$. But then $r_{ik} \in R(F)$ by the definition of $R(F)$, so every path from c_p to c_q in B contains a point from $R(F)$. Hence $R(F)$ is an articulation set for B.

The proof is similar if F is a cutset for RG.

The next three theorems reverse the process in Theorem 3.4 by starting with articulation sets in B and getting cutsets in RG or CG.

THEOREM 3.5. *Let R_0 be an articulation set for B contained in R. Let $E(r_0)$ be the set of edges $[c_i, c_j]$ in CG that satisfy $[c_i, r, c_j]$ is a path in B only if $r \in R_0$. Then $E(R_0)$ is a cutset of CG. Furthermore R_0 is an articulation set for RG if and only if the set $R - R_0$ is disconnected in B.*

Proof. First observe that if R_0 is an articulation set for RG, then $R - R_0$ is disconnected by Theorem 3.2. Also, if $R - R_0$ is disconnected, then an easy application of Lemma 3.1 shows that R_0 is an articulation set for RG.

Now suppose that R_0 is an articulation set for B contained in R. We claim that C points are disconnected in $B - R_0$. If not, then removal of R_0 disconnects the R set into at least 2 nonempty subsets, say R_1 and R_2. Then every path in B joining a point in R_1 and a point in R_2 must pass through a point $r_0 \in R_0$. Let C_1 be the set of points in C connected to R_1 in $B - R_0$, and C_2 the set of points in C connected to R_2. Since B was originally connected, C_1 and C_2 are nonempty. Furthermore, every path in B joining a point in C_1 to a point in C_2 must pass through a point $r_0 \in R_0$. Hence, the C points are disconnected in B by removal of R_0, and so removal of the edges $E(R_0)$ from CG will disconnect CG. Hence $E(R_0)$ is a cutset for CG.

Similar reasoning leads to a proof of the following theorem.

THEOREM 3.6. *Let C_0 be an articulation set for B contained in C. Let $E(C_0)$ be the set of edges $[r_i, r_j]$ in RG that satisfy $[r_i, c, r_j]$ is a path in B only if $c \in C_0$. Then $E(C_0)$ is a cutset of RG. Furthermore, C_0 is an articulation set for CG if, and only if, the set $C - C_0$ is disconnected in B.*

THEOREM 3.7. *Suppose A_0 is an articulation set for B, $A_0 = R_0 \cup C_0$, $R_0 \neq \varnothing$, $C_0 \neq \varnothing$, and R_0 and C_0 are not articulation sets for B. Then the set of edges $E(R_0) \cap (CG - \langle C_0 \rangle)$ is a cutset for the subgraph $CG - \langle C_0 \rangle$ of CG. Similarly, the set of edges $E(C_0) \cap (RG - \langle R_0 \rangle)$ is a cutset for $RG - \langle R_0 \rangle$. Moreover, R_0 is not an articulation set of RG and C_0 is not an articulation set of CG.*

Proof. We claim that removal of A_0 disconnects $C - C_0$ and $R - R_0$ in B. One possibility is that $B - A_0$ is disconnected into R_1 and C_1. This means that any path in B from C_1 to R_1 must go through R_0 and C_0. But then both R_0 and C_0 are articulation sets. Suppose A_0 disconnects R but not C. Then in $B - A_0$ we would have R disconnected into R_1 and R_2, and one of these, say R_1, connected to $C_1 = C - C_0$. But then in B, any path from C_1 to R_2 must go through R_0 and C_0 and so again both R_0 and C_0 are articulation sets.

Hence removal of A_0 from B disconnects $C - C_0$ and $R - R_0$. Therefore, removal of the edges $E(R_0) \cap (CG - \langle C_0 \rangle)$ will disconnect $CG - \langle C_0 \rangle$, and removal of the edges $E(C_0) \cap (RG - \langle R_0 \rangle)$ will disconnect $RG - \langle R_0 \rangle$.

Finally, the last statement in the theorem is an immediate consequence of Theorem 3.2.

Let us now examine tearing and show why we need more than just any articulation set to capture computer assisted analysis functions. Define $K(G)$ as the connectivity of a graph, G—that is, the minimum number of vertices whose removal increases the number of components of G. Our interest is primarily when the fundamental bigraph is connected, so $K(B) \geqq 1$. Equality holds if and only if there is an articulation vertex.

Harary [13] showed that when G is connected, $K(G)$ cannot exceed the minimum degree: $K(G) \leqq d_1$. His proof, however, reveals a difficulty. Choose a vertex of minimum degree; its adjacent vertices comprise an articulation set because their removal isolates the vertex. For the fundamental bigraph an isolated vertex corresponds to a null row or column, and this does not capture our intent. We would be more interested in finding an articulation set where the disconnected graph contains no isolated vertices.

Harary also proved that for any graph (possibly not connected), its connectivity is bounded by the average degree. This does not overcome our difficulty, but the result bears further analysis because of its ability to explain some "empirical facts". Let the matrix have m rows, n columns and z nonzeros. Thus B has $m + n$ vertices and z edges. Harary's bound is:

$$K(B) \leqq \frac{2z}{m+n}.$$

A "reasonable rule" for large matrices is that the average number of nonzeros per column is bounded by a constant. That is,

$$z \leqq cn.$$

In practice, a realistic value of c is 4 ($c = 2$ for network linear programs), and c greater than 7 is unrealistic. Thus, Harary's bound implies:

$$K(B) \leqq 2c,$$

and generally only about 8 ($c = 4$) rows and columns need to be removed to disconnect B. This agrees remarkably with empirical evidence, so another proof may reveal properties such as that no isolated vertices exist in the disconnected graph.

The following theorem gives the relationships between the connectivities of the various graphs.

THEOREM 3.8. 1) *If CG is not complete, then $K(CG) \geqq K(B)$.*

2) *If RG is not complete, then $K(RG \geqq K(B)$.*

3) $K(B) \leqq \min(K(CG)+1, K(RG)+1)$.

4) *If B is h-connected, then CG and RG are either h-connected or complete.*

Proof. 1) and 2) follow from Theorem 3.2 since any articulation set for RG or CG is an articulation set for B.

If CG is complete, then $K(CG) = n - 1$. However, removing all n column points from B disconnects B, so $K(B) \leqq n = K(CG) + 1$. Similarly, if RG is complete, $K(B) \leqq m = K(RG) + 1$. Combining this with 1) and 2) completes the proof of 3).

Finally, recall that B is h-connected if $K(B) \geqq h$. Hence, 4) follows from 1) and 2).

4. Conclusions. The main results we have obtained are Theorem 2.2, our estimate on the number of edges in CG and RG, Theorem 2.10, and our various results on articulation sets. Theorem 2.2 shows that, under mild restrictions, CG, RG and B have the same connectivity. This is only one of a number of basic properties these graphs have in common. In subsequent publications we shall derive other common properties. Theorem 2.10 establishes the basic relationships between the digraphs. Our estimates on the number of edges in CG and RG point the way to problems requiring further study whose solution will, hopefully, lead to criteria for deciding which of the graphs should be used to generate efficient algorithms for solving matrix structure problems. Finally, as we shall show in a subsequent publication, the results on articulation sets can be used to determine various possible structural forms of the rectangular matrix M.

These results can be used in computer assisted analysis and linear programming to shed light on the following problems. First we can establish a framework to render computer assistance to analysts in tracing the cause of infeasibility in LP problems so that they may debug their models. Second, we can use our results to investigate the problem of model reduction and simplification. Third, it appears that our results may be useful in helping to identify embedded structures such as networks or physical flows. Finally, we expect that we can use some of our results to help investigate sensitivity problems in linear programs.

REFERENCES

[1] L. B. BEINEKE AND F. HARARY, *The connectivity function of a graph*, Mathematika, 14 (1967), pp. 197–202.

[2] J. A. BONDY, *Properties of graphs with constraints on degrees*, Studia Sci. Math. Hung., 4 (1969), pp. 473–475.

[3] I. S. DUFF, *On algorithms for obtaining a maximal transversal*, IEEE Trans. Math. Software, to appear.

[4] A. L. DULMAZE AND N. S. MENDELSOHN, *Graphs and matrices*, in Graph Theory and Theoretical Physics, F. Harary, ed., Academic Press, New York, 1967, pp. 167–227.

[5] G. EXOO AND F. HARARY, *Step graphs*, J. Comb. Inform System Sci., to appear.

[6] H. J. GREENBERG, *Measuring complementarity and qualitative determinacy in matricial forms*, Proc. U.S. Dept. of Energy Symposium on Computer-Assisted Analysis and Model Simplification, Academic Press, New York, to appear.

[7] ———, *A new approach to analyze information contained in a model*, Proc. NBS Workshop on Validation/Assessment of Energy Models, 1979.

[8] ———, *The scope of computer-assisted analysis and model simplification*, Proc. U.S. Dept. of Energy Symposium on Computer-Assisted Analysis and Model Simplification, Academic Press, New York, to appear.

[9] H. J. GREENBERG, J. R. LUNDGREN AND J. MAYBEE, *Graph-theoretic foundations of computer-assisted analysis*, Proc. U.S. Dept. of Energy Symposium on Computer-Assisted Analysis and Model Simplification, Academic Press, New York, to appear.
[10] ———, *Structural analysis of linear programs*, to appear.
[11] ———, *Structural relationships between rectangular matrices and associated graphs*, in preparation.
[12] ———, *Adjacency matrices for graphs and digraphs associated with rectangular matrices*, submitted.
[13] F. HARARY, *The maximum connectivity of a graph*, Proc. Nat. Acad. Sci., 48 (1962), pp. 1142–1146.
[14] F. HARARY, R. NORMAN AND D. CARTWRIGHT, *Structural Models: An Introduction to the Theory of Directed Graphs*, John Wiley, New York, 1965.
[15] E. J. HENLEY AND R. A. WILLIAMS, *Graph Theory in Modern Engineering*, Academic Press, New York, 1973.
[16] J. MAYBEE AND J. QUIRK, *Qualitative problems in matrix theory*, SIAM Rev., 11 (1969), pp. 30–51.
[17] R. P. TEWARSON, *Row-column permutation of sparse matrices*, Comput. J., 10 (1967), pp. 300–305.
[18] A. W. TUCKER, *Combinatorial theory underlying linear programs*, in Recent Advances in Mathematical Programming, R. E. Graves and P. Wolfe eds., McGraw-Hill, New York, 1963, pp. 1–16.
[19] R. L. WEIL AND P. C. KETTLER, *Rearranging matrices to blockangular form for decomposition (and other) algorithms*, Management Sci., 18 (1971), pp. 98–108.
[20] I. S. DUFF, *A survey of sparse matrix research*, Proc. IEEE, 65 (1977), pp. 500–535.
[21] J. R. BUNCH AND D. J. ROSE, eds., *Sparse Matrix Computations*, Academic Press, New York, 1976.
[22] I. S. DUFF AND G. W. STEWART, eds., *Sparse Matrix Symposium* 1978, Academic Press, New York, 1979.
[23] Å. BJÖRCK, R. J. PLEMONS AND H. SCHNEIDER, eds., *Large Scale Matrix Problems*, North-Holland, New York, 1981.
[24] A. GEORGE AND J. LIU, *Computer Solution to Large Positive Definite Systems of Linear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 3, September 1981

0196-5212/81/0203-0003 $01.00/0

COVERING REGIONS WITH SQUARES*

MICHAEL O. ALBERTSON† AND CLAIRE J. O'KEEFE†

Abstract. A unit square in R^2 whose corners are integer lattice points is called a block. A board consists of a finite set of blocks. Given a board B, its graph $G(B)$ has vertices corresponding with the blocks of B, and two vertices of $G(B)$ are joined by an edge provided the corresponding blocks are contained in a square subset of B. If B is simply connected, then $G(B)$ is perfect.

A unit square in R^2 whose corners are integer lattice points is called a *block*. A finite set of blocks is called a *board*. A board B is said to be *linearly convex* if whenever two blocks of B are in the same row or column then every block between them is also in B. We reserve *rectangle* (*square*) to mean a rectangular (square) subset of the blocks of a given board. Dually we reserve *anti-rectangle* (*anti-square*) to mean a subset of the blocks of a board no two of which are contained in the same rectangle (square). Chaiken, Kleitman, Saks and Shearer [4] proved the following duality theorem about boards. If B is a linearly convex board, then the minimum number of rectangles whose union contains all the blocks of B equals the maximum number of blocks in an anti-rectangle. They also exhibit a board according to Chung which demonstrates the necessity of the convexity hypothesis. Somewhat surprisingly the dual of the above theorem is false. Specifically, Boucher has constructed a linearly convex board whose largest rectangle contains 144 blocks yet the board is not the union of 144 anti-rectangles [2].

The purpose of this paper is to establish the two dual assertions that follow.

THEOREM 1. *If B is a simply connected board, then*

(i) *The maximum number of blocks in any square of B equals the minimum number of anti-squares whose union is B; and*

(ii) *The maximum number of blocks in any anti-square of B equals the minimum number of squares whose union is B.*

For convenience, we define the graph of a board $G = G(B)$. The vertices of G correspond with the blocks of B while two vertices of G are joined by an edge if the corresponding blocks of B are contained in a square. We adopt the notation $\chi(G)$ for the chromatic number, $\omega(G)$ for the cardinality of the maximum clique, $\alpha(G)$ for the independence number, and $\theta(G)$ for the clique covering number (for definitions see [1]). It is an immediate consequence of the definitions that $\chi(G) \geqq \omega(G)$ and $\theta(G) \geqq \alpha(G)$ for any graph. The investigation of the conditions which force equality in either of the above inequalities has been a major interest of combinatorists [1], [5]. The most important result is the perfect graph theorem [6].

PERFECT GRAPH THEOREM. *If for each induced subgraph H of G, $\chi(H) = \omega(H)$, then for each such H, $\theta(H) = \alpha(H)$, and conversely.*

The proof of Theorem 1 will be accomplished by showing that the graph of a simply connected board is perfect, i.e., that it satisfies the hypotheses of the perfect graph theorem. We begin with a closer look at our board B. A block of B is said to be a *boundary block* if it intersects the boundary of B in at least one edge. If S is a square in B, the block b in S is said to be a *border block* of S if either b shares an edge with a block of $B - S$ or b is a boundary block. A *knob* of B is a $1 \times p$ rectangle three sides of which are contained in the boundary of B. A square S is said to be *disconnecting* if the interior of $B - S$ is not connected.

* Received by the editors September 26, 1980, and in revised form November 20, 1980.

† Smith College, Northampton, Massachusetts, 01063.

LEMMA 1. *Let S be a maximal square in a simply connected board B. If S is not disconnecting, then there exists a pair of opposite border lines of S each of which contains a boundary block.*

Proof. By contradiction. If S contains no pair of opposite border lines such that each line of the pair contains a boundary block of B, then there exists a pair of adjacent border lines in S each of which consists of non-boundary blocks of B. See Fig. 1. The

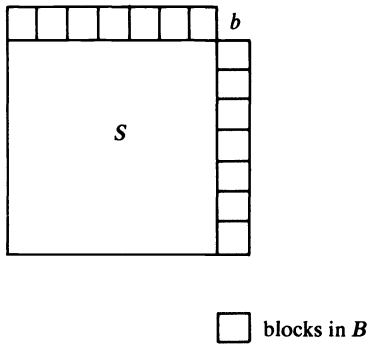


FIG. 1

maximality of S implies that b is not in B. If b is not in B, S is not disconnecting, and B is simply connected, then every other block which meets the boundary of S (even at a single point) must be in B. Hence S is not maximal. □

LEMMA 2. *Let S be a maximal square in a simply connected board B. If S is not a disconnecting square of B then S contains a knob of B.*

Proof. By Lemma 1 we may assume that b_1 and b_2 are boundary blocks contained in opposite border lines of S. See Fig. 2. It is clear that either S is disconnecting or one of the paths of border blocks from b_1 to b_2 contains a knob. □

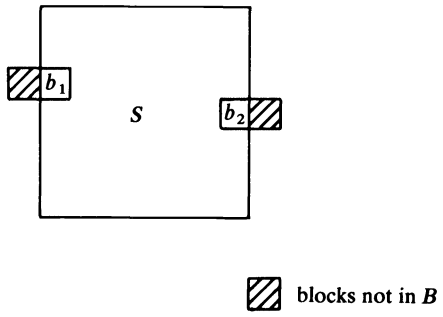


FIG. 2

As previously stated, we will show that if H is an induced subgraph of $G(B)$ then $\chi(H) = \omega(H)$. Let B' be the subset of B whose blocks correspond with the vertices of H. If S is a square of B then we will consider $S' = S \cap B'$ a "square" of B' and call B' an induced subboard of B.

Proof of Theorem 1. The proof of Theorem 1 will be by induction on the number of blocks in B. We assume the theorem holds for all simply connected boards of no more

than $(m-1)$ blocks. Let B be a simply connected board with m blocks and B' an induced subboard of B. Suppose S is a maximum square in B. By Lemma 2 there are two cases to consider.

Case (i) S is a disconnecting square of B. Let $C_1, \cdots, C_j$ denote the components of $B-S$. Two blocks are in a square of B if and only if they are in a square of $C_i \cup S$ for some i $(1 \leqq i \leqq j)$. Hence the induced subgraph with vertex set $B' \cap (C_i \cup S)$ is the same independent of whether the original graph is $G(B)$ or $G(C_i \cup S)$. Call this *induced subgraph* H_i $(1 \leqq i \leqq j)$. If $\omega(H) = N$, then $\omega(H_i) \leqq N (1 \leqq i \leqq j)$. Since $C_i \cup S$ is simply connected and smaller than B, $\chi(H_i) \leqq N (1 \leqq i \leqq j)$. Thus the blocks of $B' \cap (C_i \cup S)$ can be N-colored, no color appearing more than once in any square of $B (1 \leqq i \leqq j)$. For i $(2 \leqq i \leqq j)$ permute the colors in $(C_i \cup S) \cap B'$ so that there is agreement with the coloring of $(C_1 \cup S) \cap B'$. Since blocks in different C_i's cannot be in the same "square" in B', we can produce a coloring of B' from the coloring of the components.

Case (ii) S contains a knob K. Assume K is a row. Since S is a square there exist at least $(|K|-1)$ blocks in the column below each block of K. If at least $|K|$ blocks of B lie in the column below each block of K, then two blocks, neither in K, are in a square of B if and only if they are in a square of $B-K$. If this occurs set $L = K$. On the other hand if exactly $(|K|-1)$ blocks of B lie under b_1, a block of K, there exists a boundary block b_2 in the $(|K|-1)$st row under b_1. See Fig. 3. Since S is not disconnecting, one of the paths

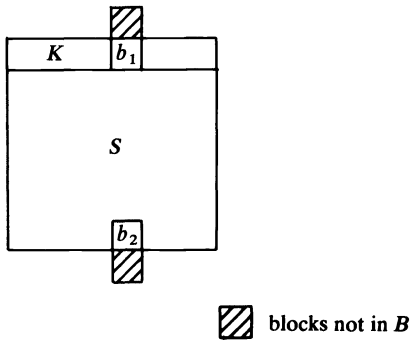


FIG. 3

of border blocks of S from b_1 to b_2 must consist entirely of boundary blocks. Thus there exist two adjacent border lines of S, say K and K^*, both of which are knobs. Set $L = K \cup K^*$. As before two blocks, neither in L, are in a square of B if and only if they are in a square of $B-K$. Let $L' = L \cap B'$, $\omega(G(B'-L')) = N$ and $\omega(G(B')) = N+\lambda\,(0 \leqq \lambda \leqq |L'|)$. By induction the blocks of $B'-L'$ can be N-colored. This coloring can be transferred to B'. By the choice of L no two blocks in a square of B are colored the same. Only the blocks of L' remain uncolored. The blocks of L' are in a square with $|S'|-|L'|$ colored blocks. The number of available colors is $N+\lambda-(|S'|-|L')$ while the number of blocks needing colors is $|L'|$. As $|S'| \leqq N+\lambda$ there are enough colors.

The above argument shows that if B is any simply connected board then $G(B)$ is perfect. Assertion (i) of Theorem 1 is immediate. By the perfect graph theorem, $\alpha(G(B)) = \theta(G(B))$. An independent set of vertices in $G(B)$ corresponds with an anti-square. A clique in $G(B)$ corresponds with a set of blocks in B, each pair contained in a square. In order to prove Assertion (ii) of Theorem 1, it is necessary to show that such a set of blocks is entirely contained in a single square. Given a set of blocks in B which corresponds with a clique in $G(B)$, it suffices to show that there is a single square in B containing a leftmost, a rightmost, a topmost, and a bottommost block in the set. It

is straightforward to check that the union of the squares which contain pairs of these extreme blocks contains a square containing all of the extreme blocks. □

The above argument is the only proof we know that $\alpha(G(B)) = \theta(G(B))$. In contrast, there are a number of alternative proofs which show $\chi(G(B)) = \omega(G(B))$. The slickest, due to Boucher [3], places no restrictions on B and proceeds as follows. Fix any coloring of any maximum square and tile the plane. Any two blocks receiving the same color in such a tiling are too far apart to be in the same square. One nice feature of this argument is that it generalizes to higher dimensions. Jim Shearer reports that both the rectangle and square 3-dimensional versions of Assertion (ii) are false. He also supplies the board in Fig. 4, which demonstrates the necessity of our hypothesis of simple connectivity [7].

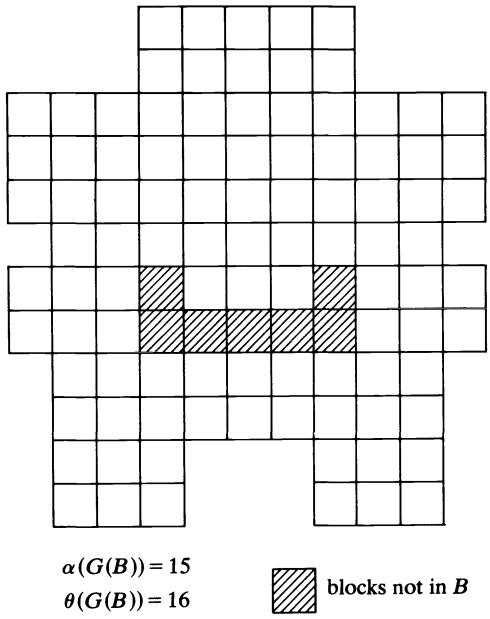


FIG. 4. *Shearer's board.*

REFERENCES

[1] CLAUDE BERGE, *Graphs and Hypergraphs*, North-Holland, Amsterdam, 1973.
[2] ANDY BOUCHER, *It's hard to color antirectangles*, this Journal, to appear.
[3] ———, personal communication.
[4] S. CHAIKEN, D. KLEITMAN, M. SAKS and J. SHEARER, *Covering regions with rectangles*, this Journal, 2 (1981), to appear.
[5] MARTIN GOLUMBIC, *Algorithmic Graph Theory and Perfect Graphs*. Academic Press, NY 1980.
[6] LAZLO LOVASZ, *A characterization of perfect graphs*. J. Combinatorial Theory (B), 13 (1972), pp. 95–98.
[7] J. SHEARER, personal communication.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 3, September 1981

0196-5212/81/0203-0004 $01.00/0

SOME RESULTS ON POLYHEDRA OF SEMIGROUP PROBLEMS*

JULIAN ARÁOZ† AND ELLIS L. JOHNSON‡

Abstract. For general additive systems we study the convex hull of solutions and its properties such as its recession cone, vertices and facets and whether it is closed. These properties depend upon various assumptions on the additive system, such as associativity (semigroups), commutativity, solvability, and generation of infeasible elements. Examples are given to illustrate the subadditive characterization of facets and to illustrate the variety of polyhedra which can arise depending upon the properties of the additive system.

1. Introduction.

Problem 1.1. *Integer programming.* The motivating problem for this work is the pure integer program problem:

$$x_j \geqq 0 \text{ and integer}, \qquad j = 1, \cdots, n,$$

$$\sum_{j=1}^{n} a_{ij}x_j = b_i, \qquad i = 1, \cdots, m,$$

$$\text{minimize } z = \sum_{j=1}^{n} c_j x_j.$$

This problem is referred to as "pure" because all of the x_j are required to take on integer values.

The practical importance of solving this problem has long been recognized. The idea of converting the problem to a linear program has also had appeal for some time. In principle, there are inequalities

$$\sum_{j=1}^{n} \alpha_{kj}x_j \geqq \beta_k, \qquad k = 1, \cdots, K$$

whose solution set is the convex hull of nonnegative integer solutions to the original system: $x_j \geqq 0$ and $\sum a_{ij}x_j = b_i$. However, finding these inequalities is a difficult, if not impossible, task in practice.

In this work, we look at the convex hull of solutions to a class of problems which includes (properly) the pure integer problem, provided bounds can be placed on the variables. In particular, pure 0-1 problems are included. However, our approach is via what is called *master problems.* We return to explain the general approach after discussing some special cases and prior work.

Problem 1.2. *Gomory's group problem.* There are several ways to derive the group problem (see [2, § 1.A]), but the simplest is to relax the equalities in the integer problem to congruences modulo 1. For rational entries a_{ij}, there are only a finite number of column vectors $\sum a_{ij}x_j$ which can be generated by integer x_j. An upper bound is the product of the least common denominators for each row.

A *master group problem* is a group problem where every column $\sum a_{ij}x_j$ which can be generated by integer x_j is already a column $A^k = (a_{1k}, a_{2k}, \cdots, a_{mk})^T$ of A for some k. These group problems have columns corresponding to nonzero elements of finite Abelian groups. Such groups are well known to be isomorphic to direct products of

* Received by the editors June 9, 1980, and in final form November 12, 1980.
† Universidad Simon Bolivar, Caracas, Venezuela.
† IBM T. J. Watson Research Center, Yorktown Heights, New York 10598.

cyclic groups, and the convex hulls of solutions have been generated [2] for groups up to size 11 (see also [3] for some corrections). These convex hulls of solutions x to group problems are what Gomory calls *corner polyhedra.*

One of Gomory's main contributions is a subadditive characterization of facets [2, Thm. 18] of corner polyhedra. That theorem has been carried over to our more general problem [4].

Gomory also shows [2, Thm. 9] that the recession cone (see § 7.1) of the convex hull of solutions is the nonnegative orthant. In addition, he shows that the vertices satisfy an irreducibility condition [2, Thm. 2], and that they are, therefore, bounded; specifically, the product $\prod(1+x_j) \leqq |G|$, where $|G|$ is the order of the group involved. He does not mention that the convex hull is closed, but that result easily follows. These three questions, asked about more general problems, are the main topic of this paper.

Problem 1.3. *Aráoz's semigroup problems.* Aráoz [1] considered semigroup problems. A special case is the *covering problem*, the integer program where all a_{ij} are nonnegative integers, b_i is a positive integer, and the restrictions are

$$x_j \geqq 0 \text{ and integer}, \qquad j = 1, \cdots, n,$$

$$\sum_{j=1}^{n} a_{ij}x_j \geqq b_i.$$

This can be viewed as a semigroup problem by defining addition $\hat{+}$ on two columns as

$$\begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix} \hat{+} \begin{pmatrix} a_{1k} \\ \vdots \\ a_{mk} \end{pmatrix} = \begin{pmatrix} \begin{cases} a_{1j}+a_{1k} & \text{if} \quad a_{1j}+a_{1k} \leqq b_1, \\ b_1 & \text{otherwise,} \end{cases} \\ \vdots \\ \begin{cases} a_{mj}+a_{mk} & \text{if} \quad a_{mj}+a_{mk} \leqq b_m, \\ b_m & \text{otherwise.} \end{cases} \end{pmatrix}.$$

Master problems are, here, problems with all integer column vectors $(\alpha_1, \cdots, \alpha_m)^T$ having $0 \leqq \alpha_i \leqq b_i$, $i = 1, \cdots, m$, present as columns of A.

For this problem, the convex hull of solutions is closed, and Aráoz [1] showed that the recession cone is R^n_+.

He also considered packing problems:

$$x_j \geqq 0 \text{ and integer}, \qquad j = 1, \cdots, n,$$

$$\sum_{j=1}^{n} a_{ij}x_j \leqq b_i.$$

In this case, the convex hull of solutions is a bounded polyhedron (i.e., a polytope), so the recession cone is just the origin. We consider this problem to be a semigroup problem by defining addition $\hat{+}$ by

$$\begin{pmatrix} a_{ij} \\ \vdots \\ a_{mj} \end{pmatrix} \hat{+} \begin{pmatrix} a_{ik} \\ \vdots \\ a_{mk} \end{pmatrix} = \begin{cases} \begin{pmatrix} a_{ij}+a_{ik} \\ \vdots \\ a_{mj}+a_{mk} \end{pmatrix} & \text{if all } a_{ij}+a_{ik} \leqq b_i, \\ \infty & \text{otherwise,} \end{cases}$$

where ∞ is a symbol used here to denote the infeasible element (see Assumption 4.3).

For Abelian semigroup problems, without ∞, Aráoz showed when the recession cone is R^n_+. Our Theorem 7.4 is an extension of that result.

For these problems, we unify the packing and covering problems (see also [4]) by allowing an infeasible element ∞. For Abelian semigroups, we answer completely the recession cone and closure questions. We also extend Gomory's notion of irreducible solution vectors in order to bound the vertices.

Problem 1.4. *Additive systems.* We generalize the semigroup problems by considering nonAbelian groups and semigroups. In addition, we consider non-associative systems. When there is no ∞, we can still answer all three questions: is the convex hull closed? (yes); what is the recession cone? (R_+^n); and what bound can be placed on the number of vertices? (§ 6). When ∞ is present, we sometimes answer the question and sometimes give counterexamples.

Problem 1.5. *Master problems.* By a master problem, we mean that there is a variable $t(g)$ for every element $g \in S$, $g \neq \hat{O}$ and $g \neq \infty$. A particular problem with only a subset T of S present has a convex hull which is the intersection of the convex hull of solutions to its master problem with bounding faces $t(g) = 0$, if $g \notin T$. We get its vertices by taking all vertices for the master problem having $t(g) = 0$ for $g \notin T$. Its facets are among the inequalities one gets by deleting $t(g)$, $g \notin T$, from the facets of the master problem convex hull (see Gomory [2, Thms. 12 and 13]).

2. Expressions in additive systems. We follow here the development in [4].

DEFINITION 2.1. The pair $(S, \hat{+})$ is an *additive system* if

$$g \hat{+} h \in S \quad \text{for all } g, h \in S, \qquad (\textit{closure}).$$

Thus, we allow very general addition but consider various restrictions which will be placed on $(S, \hat{+})$. The most commonly used restriction is *associativity*:

$$g \hat{+} (h \hat{+} k) = (g \hat{+} h) \hat{+} k \quad \text{for all } g, h, k \in S, \qquad (\textit{associativity}).$$

If $(S, \hat{+})$ is associative, then it is called a *semigroup.* Another property is *commutativity*:

$$g \hat{+} h = h \hat{+} g \quad \text{for all } g, h \in S \qquad (\textit{commutativity}).$$

If commutativity holds, we refer to $(S, \hat{+})$ as being *Abelian.*

DEFINITION 2.2 (Expressions). For an additive system $(S, \hat{+})$, an *expression* E of $(S, \hat{+})$ is defined recursively by:

(i) (g) is an expression, for all $g \in S$;

(ii) $(E_1 \hat{+} E_2)$ is an expression, whenever E_1 and E_2 are expressions.

An expression (g) is called a *primitive expression.* When an expression $E = (E_1 \hat{+} E_2)$, as in (ii), we call E_1 and E_2 *subexpressions* of E, and any subexpression of E_1 or E_2 is also a subexpression of E. *Primitive subexpressions* of E are those subexpressions of E which are primitive expressions of the form (g), $g \in S$.

DEFINITION 2.3 (Evaluation). An expression is to be thought of as simply a string of symbols, $(\cdot\,, \cdot)$, $\hat{+}$ and g for $g \in S$. The *evaluation* of an expression E is a function γ from expressions to S defined recursively by:

(i) $\gamma(E) = g$, if $E = (g)$, $g \in S$;

(ii) $\gamma(E) = \gamma(E_1) \hat{+} \gamma(E_2)$ if $E = E_1 \hat{+} E_2$.

To *evaluate* E means to find $\gamma(E)$, which can be done recursively by the definition.

LEMMA 2.4 (Substitution lemma). *If E_1 is a subexpression of E with $\gamma(E_1) = g$, then (g) can be put in place of E_1 in E without changing the evaluation of E.*

The proof of this lemma is more or less clear from the inductive definitions of expressions and evaluations. However, it is a frequently used fact which is convenient to explicitly state.

DEFINITION 2.5 (Incidence vector). A vector $(t(g), g \in S)$ is the *incidence vector of an expression* E if $t(g)$ is equal to the number of times (g) appears as a primitive expression of E.

DEFINITION 2.6 (t represents g). For an incidence vector t, define *t represents g* to mean that there is some expression E for which t is the incidence vector of E, and $\gamma(E) = g$.

3. The convex hull $\mathcal{H}(S, b)$.

DEFINITION 3.1 (Solution vector). Fix some element $b \in S$, for $(S, \hat{+})$ an additive set, and call b the *right-hand side*. An expression E is a *solution expression* if $\gamma(E) = b$. The incidence vector t of an expression E is a *solution vector* if E is a solution expression. That is, t is a solution vector if it represents b (see Definition 2.6).

DEFINITION 3.2 (Convex hull $\mathcal{H}(S, b)$). Define the *convex hull* of solutions to be

$$\mathcal{H}(S, b) = \text{conv}\,\{(t(g), g \in S) | t \text{ is a solution vector}\}.$$

Our problem is to determine properties of $\mathcal{H}(S, b)$ such as its vertices, facets and recession cone.

DEFINITION 3.3 (Master problem). We only consider here what are called *master problems* in that $\mathcal{H}(S, b)$ is taken over all $g \in S$. A *subproblem* would be given if we restricted the primitive expressions (g) to have $g \in S'$ for some subset $S' \subseteq S$. However, as in [1] and [2], one can get polyhedra for subproblems from those of master problems by projecting onto a face ($t(g) = 0$ for $g \notin S'$). See also our discussion in Problem 1.5.

4. Simplifications and reductions. The arguments given here will lead to assumptions on $(S, \hat{+})$, without loss of generality. These assumptions are important in order to avoid confusion in subsequent sections.

Assumption 4.1 (Zero $\hat{O}$). We assume an element $\hat{O} \in S$ such that $\hat{O} \hat{+} g = g \hat{+} \hat{O} = g$ for all $g \in S$. If such an element were not in S, we could adjoin it to S without changing, for our purposes, the additive set $(S, \hat{+})$. There can, clearly, be only one zero in $(S, \hat{+})$.

Assumption 4.2 (The empty expression). It is convenient to assume that the empty string is also an expression whose evaluation is $\hat{O}$.

Assumption 4.3 (Infinity ∞). Every element $g \in S$ such that $t(g) = 0$ in every solution vector can be collapsed to a single element called *infinity* and denoted by ∞. It satisfies $h \hat{+} \infty = \infty \hat{+} h = \infty$ for all $h \in S$. We do not assume that there is always an $\infty \in S$; it is only present if $g \hat{+} h = \infty$ for some g and h in S, such that $s(g) > 0$ in some solution s and $t(h) > 0$ in some solution t.

Assumption 4.4 (Feasibility assumption). Introduction of ∞ leads to the assumption that for $g \in S$, $g \neq \hat{O}$ or ∞, there is some solution vector t having $t(g) > 0$.

Assumption 4.5 (Nonzero b). We assume that $b \neq \hat{O}$ and $b \neq \infty$.

Assumption 4.6 (Nonzero subexpressions). In any solution expression E, we delete any primitive subexpression $(\hat{O})$ without changing $\gamma(E)$.

Assumption 4.7 (Deletion of $t(\hat{O})$ and $t(\infty)$). With Assumptions 4.3 and 4.6, $t(\hat{O}) = 0$ and $t(\infty) = 0$ in any solution vector, and we delete both $t(\hat{O})$ and $t(\infty)$ from t so that solution vectors have the form $t = (t(g), g \in S - \{\hat{O}, \infty\})$.

DEFINITION 4.8. The *finite elements* of S are denoted by S_f; that is,

$$S_f = S - \{\infty\}.$$

The *proper elements* of S are $g \in S_p$, where

$$S_p = S - \{\hat{O}, \infty\}.$$

Then our incidence vectors are of the form

$$t = (t(g),\ g \in S_p).$$

5. Generators and loops.

DEFINITION 5.1 (Subsystem). Define a *subsystem* $(T, \hat{+})$ of $(S, \hat{+})$ to be a subset T of S with the same addition as in $(S, \hat{+})$ and such that $g \hat{+} h \in T$ for all $g, h \in T$.

For $G \subseteq S$, define the *system generated by* G to be the set

$$\{h | h = \gamma(E) \text{ all expressions } E \text{ made from primitive expressions } (g),\ g \in G\},$$

together with the $\hat{+}$ from $(S, \hat{+})$. The fact that the system generated by G is a subsystem of $(S, \hat{+})$ follows from the inductive definition of expressions and evaluations:

$$\text{If } h_1 = \gamma(E_1) \text{ and } h_2 = \gamma(E_2), \text{ then } h_1 \hat{+} h_2 = \gamma((E_1 \hat{+} E_2)).$$

Remark 5.2 (Loops in semigroups). An especially interesting case is the subsystem generated by a single element. When $(S, \hat{+})$ is a semigroup, i.e., associativity holds, then for $g \in S$, and k a nonnegative integer, $kg \in S$ is well defined as $\gamma(g \hat{+} g \hat{+} \cdots \hat{+} g)$ where g is taken k times and parentheses are not needed because of associativity. Consider the sequence

$$\hat{O},\quad g,\quad 2g,\quad 3g,\quad \cdots,\quad kg,\quad \cdots.$$

Since S is a finite set, there can only be a finite number of different elements. Let

$$h_0 = mg$$

be the first occurrence of any element appearing for the second time in the sequence. Define the *order of* g to be m. Since h_0 appears earlier in the sequence, $h_0 = kg$ for some $k < m$. The sequence of distinct semigroup elements

$$h_0 = kg,\quad h_1 = (k+1)g,\quad \cdots,\quad h_{m-k-1} = (m-1)g$$

is the same as

$$mg,\quad (m+1)g,\quad \cdots,\quad (2m\text{-}k\text{-}1)g,$$

and, in fact, repeats itself indefinitely in the sequence $\hat{O}, g, 2g, \cdots$. The sequence

$$h_0, h_1,\quad \cdots,\quad h_{m-k-1}$$

is called the *loop of* g and $l = m - k$ is called the *loop order of* g. Define g to be a *loop element* of $(S, \hat{+})$ if g belongs to its loop, i.e., k is either 0 or 1.

DEFINITION 5.3 $(g \to \infty)$. If ∞ is equal to any kg then the loop of g must be the sequence of one element, namely ∞. Thus, the loop order is 1. In this case, we say g *goes to* ∞ and write $g \to \infty$. If g does not go to ∞, then we write $g \not\to \infty$.

LEMMA 5.4 (Abelian subsemigroup lemma). *If* $(S, \hat{+})$ *is a semigroup, then the subsemigroup*

$$G = \{kg | k \text{ integer and } k \geqq 0\},$$

is an Abelian subsemigroup of $(S, \hat{+})$ *for all* $g \in S$.

Proof. What we need to show is that

$$(kg) \hat{+} (lg) = (lg) \hat{+} (kg)$$

for all integers $k, l \geqq 0$. This follows from $(kg) \hat{+} (lg) = (k+l)g$, since associativity allows us to remove the parentheses. As a corollary, we have the following fact.

COROLLARY 5.5. *For any two elements h_1 and h_2 in the loop of g, $h_1 \hat{+} h_2 = h_2 \hat{+} h_1$, provided $(S, \hat{+})$ is a semigroup.*

DEFINITION 5.6 (Loops in non-associative systems). We must now define

$$kg = \{h | h = \gamma(E) \text{ over all expressions } E \text{ having } k \text{ primitive expressions each of which is } (g)\}.$$

Let $0g = \{\hat{O}\}$ and $1g = \{g\}$. Because there are only a finite number of subsets of S, the sequence of sets

$$0g, \quad 1g, \quad 2g, \quad 3g, \quad \cdots, \quad kg, \cdots$$

must eventually have a first set which appeared earlier. We can now define the loop of g, the order of g, and the loop order of g as before but in terms of sets of elements rather than single elements.

DEFINITION 5.7 ($g \to \infty$). If in the loop of g there is only one set, and if it is the singleton $\{\infty\}$, then we define g *goes to* ∞ and write $g \to \infty$. Otherwise, there is some finite element in every set in the loop of g and g does not go to ∞: $g \not\to \infty$.

6. Extreme points of $\mathcal{H}(S, b)$. In the case of Abelian groups, Gomory [2] defined the notion of irreducible vector in order to prove that

$$\prod_{g \in S_p} (t(g) + 1) \leqq |S|$$

for any vertex t of $\mathcal{H}(S, b)$. We try to follow a parallel development here.

DEFINITION 6.1 (Irreducible solution vectors). A solution vector t is *reducible* if among all of the solution expressions for which t is the incidence vector, there are two expressions E_1 and E_2 with subexpressions E_3 and E_4, respectively, such that

$$\gamma(E_3) = \gamma(E_4)$$

and

$$r \neq s,$$

where r is the incidence vector of E_3 and s is the incidence vector of E_4. If a solution vector is not reducible, it is called an *irreducible solution vector.*

DEFINITION 6.2 (Extreme points or vertices). Define an *extreme point*, or a *vertex*, of a convex set C to be a point $t \in C$ such that there do not exist t^1 and $t^2 \in C$, $t^1 \neq t^2$, such that

$$t = \tfrac{1}{2}t^1 + \tfrac{1}{2}t^2.$$

It is equivalent so say that t is an extreme point of C if there exists an objective function

$$z(x) = \sum_{g \in S} c(g)x(g),$$

such that the minimum value of $z(x)$ over $x \in C$ is given uniquely by $x = t$.

THEOREM 6.3. *All extreme points of $\mathcal{H}(S, b)$ are irreducible.*

Proof. The proof follows Gomory [2]. Let t be reducible. We must show it cannot be an extreme point of $\mathcal{H}(S, b)$.

By Definition 6.1, there exist solution expressions E_1 and E_2, with subexpressions E_3 and E_4 respectively, such that $\gamma(E_3) = \gamma(E_4)$ and $r \neq s$, where r and s are the incidence vectors of E_3 and E_4 respectively.

Form expressions E_1' and E_2' from E_1 and E_2, as follows: in E_1 replace E_3 by E_4 to give E_1' and in E_2 replace E_4 by E_3 to give E_2'. By Lemma 2.4,

$$\gamma(E_1') = \gamma(E_2') = \gamma(E_1) = \gamma(E_2) = b.$$

Thus, E_1' and E_2' are solution expressions. Their incidence vectors are

$$t^1 = t - r + s, \quad \text{and} \quad t^2 = t - s + r,$$

respectively. Clearly,

$$t = \tfrac{1}{2}t^1 + \tfrac{1}{2}t^2,$$

proving that t is not a vertex of $\mathcal{H}(S, b)$.

THEOREM 6.4. *If t is a vertex of $\mathcal{H}(S, b)$, then*

$$\sum_{g \in S} t(g) \leqq 2^{|S|}.$$

Proof. If t is a vertex of $\mathcal{H}(S, b)$, then by Theorem 6.3, t is an irreducible solution vector. In particular, t is the incidence vector of a solution expression E such that any two subexpressions E_1 and E_2 having different incident vectors r and s, respectively, must have different evaluations. Even weaker, if $\sum r(g) \neq \sum s(g)$, then $\gamma(E_1) \neq \gamma(E_2)$. This statement is, in fact, all that we use about t being a vertex.

Consider how E is formed. Definition 2.1 says that E can be decomposed into

$$E = (E_1 \hat{+} E_2),$$

and each of E_1, E_2 can be similiarly decomposed unless it is primitive. Thus, E can be thought of as a binary tree with each node being a subexpression with the root being E and each end of the tree being a primitive node. Along any path from the root to an end, the subexpressions have different lengths so must have different evaluations. Hence, no such path can be longer than $|S|$. The value of $\sum t(g)$ is, in fact, the number of ends of the tree and is maximized by a complete binary tree, i.e., one have 2^d ends where d is the length of every path from the root to an end. Thus, the inequality

$$\sum_{g \in S} t(g) \leqq 2^{|S|}$$

follows.

THEOREM 6.5. *If $(S, \hat{+})$ is a semigroup, then every vertex t of $\mathcal{H}(S, b)$ satisfies*

$$\sum_{g \in S} t(g) \leqq |S|.$$

Proof. Let t be a vertex and the incidence vector of the solution expression E. By associativity, the parentheses in E can be rearranged, without changing its evaluation, so that

$$E = (\cdots ((((g_1) \hat{+} (g_2)) \hat{+} (g_3)) \hat{+} (g_4)) \hat{+} \cdots).$$

Now, E has subexpressions

$$(g_1), \quad ((g_1) \hat{+} (g_2)), \quad (((g_1) \hat{+} (g_2)) \hat{+} (g_3)), \quad \cdots,$$

of increasing length. Since t is a vertex, E must be irreducible, so each of these subexpressions must have different evaluations because they clearly have different incidence vectors, being of increasing length. Therefore, the number of primitive subexpressions in E is at most $|S|$, and the result follows.

THEOREM 6.6 *If* $(S, \hat{+})$ *is an Abelian semigroup, then every vertex of* $\mathscr{H}(S, b)$ *satisfies*

$$\prod_{g \in S} (1 + t(g)) \leqq |S|.$$

The proof of Gomory [2] for Abelian groups carries over directly to prove this theorem where group is replaced by semigroup.

7. Extreme rays of $\mathscr{H}(S, b)$.

DEFINITION 7.1 (Recession cone). The *recession cone* of a convex set C is all of the vectors r such that $x + \gamma r \in C$ for some $x \in C$ and all $\gamma \geqq 0$. When C is closed as well as convex, then for any r in the recession cone $(x + \gamma r) \in C$ for all $x \in C$. That is, the choice of $x \in C$ is irrelevant. Section 8 addresses the closure question.

THEOREM 7.2. *The recession cone of* $\mathscr{H}(S, b)$ *is the nonnegative orthant* R^d_+ *whenever* $g \not\to \infty$ *for all* $g \in S_p$.

Here $d = |S_p|$, where S_p is the set of proper elements of S as defined in Definition 4.8.

Proof. It should be clear that the recession cone of $\mathscr{H}(S, b)$ must be contained in the nonnegative orthant R^d_+ because $\mathscr{H}(S, b)$ is contained in it. Hence, we need only show the reverse inclusion.

Let $g \in S_p$. There is an $h \in S_p$ in the loop of g because, first of all, we can find an $h \neq \infty$ in the loop of g by $g \not\to \infty$. Further, if $\hat{O}$ is in the loop of g, then so is g and $g \in S_p$.

In the Abelian semigroup case, we can just say that $h \hat{+} ilg = h$ for all positive integers i, where l is the loop order of g. In general, we can say that there is some expression E_i containing (h) once and (g) il times as primitive subexpressions, and such that $\gamma(E_i) = h$.

Because $h \neq \infty$, there is some solution E containing (h) as a primitive subexpression. By Lemma 2.4, we can form solution expressions E'_i by putting E_i in place of (h). Let t be the solution vector which is the incidence vector of E. Then t^i, which is the solution vector which is the incidence vector of E'_i, satisfies

$$t^i = t + il\delta_g,$$

where

$$\delta_g(f) = \begin{cases} 1 & \text{if } f = g, \\ 0 & \text{if } f \neq g, \quad f \in S_p. \end{cases}$$

We have, thus, proven that δ_g is in the recession cone of $\mathscr{H}(S, b)$ for all $g \in S_p$.

COROLLARY 7.3. *In the case when* $(S, \hat{+})$ *is a group, even a nonAbelian group, the recession cone of* $\mathscr{H}(S, b)$ *is all of the nonnegative orthant* R^d_+.

Proof. In the group case, there can be no ∞ because ∞ has no inverse.

We have established the recession cone in the case, in particular, when there is no ∞, even if $(S, \hat{+})$ is not associative.

THEOREM 7.4. *For an Abelian semigroup* $(S, \hat{+})$, *the recession cone of* $\mathscr{H}(S, b)$ *is the cone*

$$s(g) = 0, \quad g \in S_p, \quad \text{and} \quad g \to \infty,$$

$$s(g) \geqq 0, \quad g \in S_p, \quad \text{and} \quad g \not\to \infty.$$

Proof. The proof of Theorem 7.2 suffices to show that δ_g is in the recession cone if $g \not\to \infty$. It remains to show that any other rays in the recession cone have $s(g) = 0$, if

$g \to \infty$. In order to show that fact, it suffices to show that $t(g) \leqq k$ for every solution vector t, where k is the order of g and $g \to \infty$. Thus, $(k+1)g = \infty$. If t is a solution vector, then it is the incidence vector of some solution expression E. Because $(S, \hat{+})$ is an Abelian semigroup, we can rearrange E to get another expression E' having the same incidence vector t and such that all occurrences of g are together. If there were more than k occurrences of g and $(k+1)g = \infty$, then $\gamma(E') = \infty$, because $(k+1)g$ is already ∞ and adding other elements will not change the evaluation. Thus, a contradiction is reached.

LEMMA 7.5. *For a semigroup $(S, \hat{+})$ with ∞, there exists an $h \in S_p$ such that $h \to \infty$.*

Proof. In the case where $(S, \hat{+})$ is Abelian, the proof is easy, and an even stronger result (Lemma 8.2) is shown in the next section. Thus, we think of $\hat{+}$ as being non-commutative.

Suppose $h \hat{+} g = \infty$. If any $a \in S_p$ has $g \hat{+} a \hat{+} h \neq \infty$, then

$$(g \hat{+} a \hat{+} h) \hat{+} (g \hat{+} a \hat{+} h) = (g \hat{+} a) \hat{+} (h \hat{+} g) \hat{+} (a \hat{+} h) = \infty.$$

That is, the element $(g \hat{+} a \hat{+} h) \to \infty$. Suppose, then, that for all $a \in S_p$, $g \hat{+} a \hat{+} h = \infty$.

Now, there exist elements h_l, h_r, g_l, g_r such that

$$h_l \hat{+} h \hat{+} h_r = b \quad \text{and} \quad g_l \hat{+} g \hat{+} g_r = b.$$

Therefore,

$$b \hat{+} b = g_l \hat{+} (g \hat{+} g_r \hat{+} h_l \hat{+} h) + h_r = \infty,$$

since

$$g \hat{+} g_r \hat{+} h_l \hat{+} h = \infty.$$

Thus, in this case $b \to \infty$. Therefore, there must be some element a in S_p such that $a \to \infty$, completing the proof.

THEOREM 7.6. *Let $(S, \hat{+})$ be a semigroup with ∞. Then the recession cone cannot be equal to the nonnegative orthant R_+^d.*

Proof. By Lemma 7.5, there exists $h \to \infty$, $h \in S_p$. We will show that δ_h is not a recession direction.

Since $h \neq \infty$ and $h \to \infty$, there is some positive integer k such that $kh \neq \infty$ and $(k+1)h = \infty$. There cannot be more than k consecutive h's in any solution expression. Hence, any solution vector t satisfies

$$\sum_{g \neq h} t(g) \geqq \frac{t(h)}{k} - 1,$$

and hence

$$\sum_{g \neq h} kt(g) - t(h) \geqq -k,$$

is a valid inequality for $\mathcal{H}(S, b)$. For any $x \in \mathcal{H}(S, b)$, $x + \lambda \delta_h$ would violate this inequality for λ large enough. Hence, δ_h is not a recession direction of $\mathcal{H}(S, b)$.

We do not characterize the recession cone for non-Abelian semigroups with ∞, except that it is not R_+^d, or for nonassociative systems with ∞. In those cases, we only give examples showing what might happen. See Table 1, where the notation (A1), for example, refers to Example 1 of the Appendix.

8. Closure. In this section, we address the question of when $\mathcal{H}(S, b)$ is closed. If it is closed, then it is polyhedral because it has only a finite number of facets [5].

TABLE 1
Summary of recession cone results.

Groups	Semigroups				Nonassociative	
	Abelian		Non-Abelian			
	no ∞	with ∞	no ∞	with ∞	no ∞	with ∞
R_+^d (Cor 7.3)	R_+^d (Thm 7.2)	$R_+^k, k<d$ (Thm 7.4)	R_+^d (Thm 7.2)	not $= R_+^d$ (Thm 7.6)	R_+^d (Thm 7.2)	could be R_+^d (A3, A4)
				could be $R_+^k(A_1)$		could be R_+^k(A5)
				could have other rays (A2)		could have other rays (A6)

LEMMA 8.1. *If $(S, \hat{+})$ is a semigroup such that, for each $g \in S_p$ there exists an $h \in S$ such that*

$$g \hat{+} h = b \quad or \quad h \hat{+} g = b,$$

then $b \to \infty$ if and only if b is not a loop element.

Proof. If $b \to \infty$, then clearly b is not a loop element.

Let $b \not\to \infty$. Suppose b is not a loop element. Since $b \not\to \infty$, there is some $g \neq \infty$ in the loop of b. Thus,

$$g = kb \hat{+} ilb \quad \text{for all } i \geqq 0,$$

where l is the loop order of b. We can use this notation without parentheses because $(S, \hat{+})$ is a semigroup. The order does not matter by Lemma 5.4. Thus,

$$g = g \hat{+} lb = lb \hat{+} g.$$

By assumption there is some $h \in S$ such that $g \hat{+} h = b$, or $h \hat{+} g = b$. Assume $h \hat{+} g = b$. Then, $b = h \hat{+} g = h \hat{+} g \hat{+} lb = b \hat{+} lb = (l+1)b$, contradicting the assumption that b is not a loop element.

LEMMA 8.2. *If $(S, \hat{+})$ is an Abelian semigroup, then ∞ is in S if and only if $b \to \infty$. If $b \not\to \infty$ then b is a loop element.*

Proof. The last part follows from Lemma 8.1 and commutativity. One half of the first part is easy: if $b \to \infty$ then clearly S has an ∞.

Let $b \not\to \infty$. Suppose $g \hat{+} h = \infty$ for some $g, h \in S_p$. There are some $\bar{g}, \bar{h} \in S_p$ such that $g \hat{+} \bar{g} = b$ and $h \hat{+} \bar{h} = b$.
Then,

$$b \hat{+} b = g \hat{+} \bar{g} \hat{+} h \hat{+} \bar{h} = g \hat{+} h \hat{+} \bar{g} \hat{+} \bar{h} = \infty \hat{+} \bar{g} \hat{+} \bar{h} = \infty,$$

contradicting $b \not\to \infty$. Hence, no two g, h in S_p can add to ∞, and $\infty \notin S$.

LEMMA 8.3. *If $(S, \hat{+})$ is a semigroup and if $g \not\to \infty$ and g has loop order l, then*

$$b = b \hat{+} ilg, \quad or \quad b = ilg \hat{+} b \quad for\ all\ i \geqq 0,$$

provided some element h in the loop of g has an $\bar{h} \in S_p$ such that

$$b = \bar{h} \hat{+} h \quad or \quad b = h \hat{+} \bar{h}.$$

Proof. Let $g \not\to \infty$ and let h be in the loop of g such that $b = h \hat{+} \bar{h}$. By Lemma 5.4, the subsemigroup generated by g is Abelian. Thus, $h = ilg \hat{+} h$ for all $i \geqq 0$, where l is the loop order of g. Hence, $b = h \hat{+} \bar{h} = ilg \hat{+} h \hat{+} \bar{h} = ilg \hat{+} b$, and the lemma is proven.

LEMMA 8.4. *To show that $\mathcal{H}(S, b)$ is closed, it suffices to show that for any extreme ray d of the recession cone, $\delta_b + kd$ is a solution vector for k which can be made arbitrarily large.*

Proof. Clearly, $\mathcal{H}(S, b)$ is closed if $t + kd$ can be shown to be in $\mathcal{H}(S, b)$ for all $t \in \mathcal{H}(S, b)$. It suffices to show that fact for all vertices t of $\mathcal{H}(S, b)$. We propose to show even more; namely that $t + kd$ is a solution vector for all vertices t. If t is a vertex, then it is a solution vector and, thus, the incidence vector of a solution expression E. If $\delta_b + kd$ is a solution vector, then it is the incidence vector of some solution expression E'. If we substitute E in place of (b) in E', then Lemma 2.4 says that the resulting expression is a solution expression, and it does have incidence vector $t + kd$, completing the proof.

THEOREM 8.5. *If $(S, \hat{+})$ is an Abelian semigroup with $\infty \notin S$, then the recession cone of $\mathcal{H}(S, b)$ is R_+^d, and $\mathcal{H}(S, b)$ is closed.*

Proof. The recession cone result was shown in Theorem 7.4. Since $(S, \hat{+})$ is Abelian, Lemma 8.3 shows that $b = b \hat{+} ilg$, $i \geqq 0$. Hence, $\delta_b + il\delta_g$ is a solution vector for $i \geqq 0$, and Lemma 8.4 completes the proof of the theorem since the extreme rays of R_+^d are just δ_g, $g \in S_p$.

THEOREM 8.6. *If $(S, \hat{+})$ is an Abelian semigroup with ∞, then $b \to \infty$ and $\mathcal{H}(S, b)$ is closed, but the recession cone has extreme rays equal to δ_g for all $g \not\to \infty$. Thus, the recession cone is equal to some R_+^m, $m < d$.*

Proof. Lemma 8.2 shows that $b \to \infty$. Theorem 7.4 shows the recession cone result. As in the proof of Theorem 8.5, Lemmas 8.3 and 8.4 suffice to complete the proof.

THEOREM 8.7. *Let $(S, \hat{+})$ be a non-Abelian semigroup without ∞ and such that for each $g \in S_p$ there exists an $h \in S$ such that $g \hat{+} h = b$, or $h \hat{+} g = b$. Then, $\mathcal{H}(S, b)$ is closed and the recession cone of $\mathcal{H}(S, b)$ is R_+^d.*

Proof. The proof is virtually the same as that of Theorem 8.5 except that Theorem 7.2 is used in place of Theorem 7.4.

Appendix. Several examples are given in the figures below. Some are special cases used to illustrate various points in the paper.

In the addition tables accompanying the figures, $\hat{O}$ and ∞ are not included. The elements are considered to be numbered $\hat{O} = g_0, g_1, g_2, \cdots, g_k = b$. We leave out the g and use only the subscript.

A1. Non-Abelian semigroup with ∞ and recession cone R_+^k, $k < d$.

			Vertices			*Rays*			*Facets*
1	3	3							
∞	2	∞	1	1	0	1	0	0	$t_3 \leqq 1$
∞	3	∞	0	0	1	0	1	0	$t_1 + t_3 \geqq 1$
									$t_2 + t_3 \geqq 1$

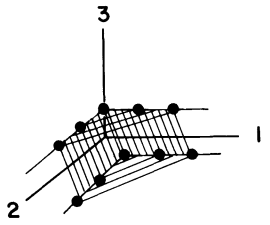


FIG. 1

A2. Non-Abelian semigroup with ∞ and ray not equal to a coordinate-direction and $\mathcal{H}(S, b)$ not closed.

	Vertices	*Rays*	*Facets*
1 2 1 2	0 0 0 1	1 0 0 0	$t_2+t_4 \geqq 1$
∞ ∞ 1 2	0 1 1 0	0 0 1 0	$-t_2+t_3 \geqq 0$
3 4 3 4		0 0 0 1	
∞ ∞ 3 4		0 1 1 0	

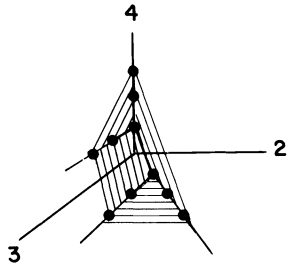


FIG. 2

A3. Non-associative system with ∞ and recession cone R_+^d.

	Vertices	*Rays*	*Facets*
1 ∞ 3	0 0 1	1 0 0	$t_3 \geqq 1$
∞ 2 3		0 1 0	
3 3 3		0 0 1	

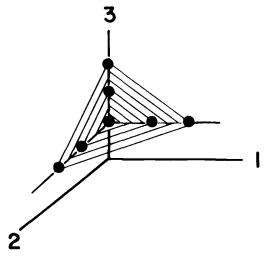


FIG. 3

A4. Non-associative system with $g \to \infty$ and recession cone R_+^d.

	Vertices	*Rays*	*Facets*
∞ 2	0 1	1 0	$t_2 \geqq 1$
2 2		0 1	

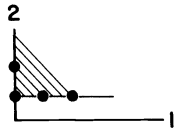


FIG. 4

A5. Non-associative system with ∞ and recession cone R_+^k, $k < d$.

	Vertices	*Rays*	*Facets*
∞ 3 ∞	0 0 1	0 1 0	$-t_1 \geqq 1$
∞ 2 3	1 1 0	0 0 1	$-t_1 + t_2 \geqq 0$
∞ 3 3			$t_1 + t_3 \geqq 1$

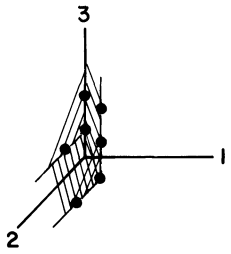


FIG. 5

A6. Non-associative system with ∞ and extreme ray not a coordinate direction.

	Vertices	*Ray*	*Facets*
∞ 0 ∞	0 0 1	0 1 1	$t_3 = 1$
∞ ∞ ∞			$t_1 - t_2 = 0$
∞ ∞ ∞			

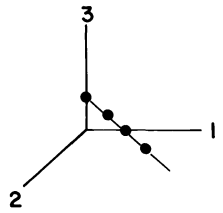


FIG. 6

A7. Non-Abelian semigroup with $\mathcal{H}(S, b)$ not closed.

	Vertices	*Rays*	*Facets*
1 2 1 2	0 1 1 0	1 0 0 0	$t_2 + t_4 \geqq 1$
1 2 1 2	0 0 0 1	0 1 0 0	$t_3 + t_4 \geqq 1$
3 4 3 4		0 0 1 0	
3 4 3 4		0 0 0 1	

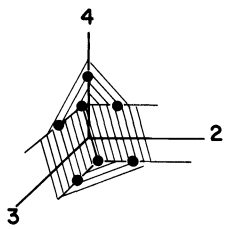


FIG. 7

A8. Non-associative system without ∞ and such that $\mathcal{H}(S, B)$ is not closed.

	Vertices	*Rays*	*Facets*
1 0	0 1	1 0	$t_2 \geqq 1$
0 0		0 1	

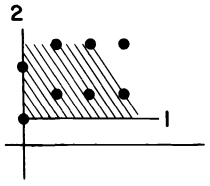


FIG. 8

A9. Non-associative system with ∞ and such that $\mathcal{H}(S, b)$ is not closed.

	Vertices	*Rays*	*Facets*
1 0	0 1	1 0	$t_2 \geqq 1$
∞ ∞		1 1	$t_1 - t_2 \geqq -1$

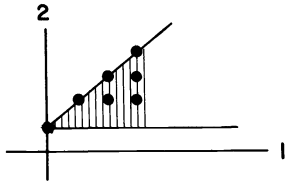


FIG. 9

A10. Non-associative system with element having no left or right complementor, but $\mathcal{H}(S, b)$ closed.

	Vertices	*Rays*	*Facets*
2 2 2	4 0 0	1 0 0	$t_1 + 2t_2 + 4t_3 \geqq 4$
2 3 3	0 2 0	0 1 0	
2 3 3	0 0 1	0 0 1	

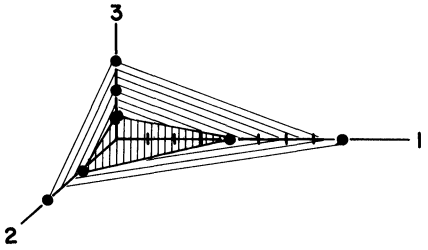


FIG. 10

REFERENCES

[1] J. ARÁOZ, *Polyhedral neopolarities*, Ph.D. thesis, Dept. of Computer Sciences and Applied Analysis, Univ. of Waterloo, Waterloo, Ontario, 1973.

[2] R. E. GOMORY, *Some polyhedra related to combinatorial problems*, Linear Algebra Appl., 2(1969), pp. 451–558.

[3] T. C. HU, *Integer Programming and Networks Flows*, Addison-Wesley, Reading, MA., 1969.

[4] E. L. JOHNSON, *On the generality of the subadditive characterization of facets*, RC 7521, IBM Research, Yorktown Heights, N.Y. 10598; Math. Oper. Res., to appear.

[5] ———, *Integer Programming: Facets, Subadditivity, and Duality for Group and Semi-group Problems*, CBMS-NSF Regional Conference Series in Applied Mathematics, 32, Society for Industrial and Applied Mathematics, Philadelphia, 1980.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 3, September 1981

0196-5212/81/0203-0005 $01.00/0

M-MATRICES WHOSE INVERSES ARE STOCHASTIC*

RONALD L. SMITH†

Abstract. This paper characterizes M-matrices whose inverses are stochastic. Such matrices can be used to model physical systems which return to equilibrium after minor disturbances. All solutions to linear systems defined by these models return to equilibrium as $t \to \infty$ at a common uniform rate.

1. Introduction. The discussion in the following paragraph is taken from Bellman [1, pp. 240–242].

Matrices whose characteristic roots have negative real parts arise in stability theory. In particular, suppose that one is interested in the behavior of a physical system in the neighborhood of an equilibrium state. A system is said to be *stable* if it returns to the equilibrium state after being subjected to small disturbances. A linear system of the form

$$\frac{dx}{dt} = Ax, \qquad x(0) = c \tag{1}$$

can often be used to study the behavior of such a system in the vicinity of the equilibrium position, which in this case is $x = 0$. A necessary and sufficient condition that the solution of (1) approach zero as $t \to \infty$ is given by the following theorem.

THEOREM A [1, p. 241]. *A necessary and sufficient condition that the solution of* (1) *regardless of the value of c, approach zero as* $t \to \infty$, *is that all characteristic roots of A have negative real parts. Consequently, we say that a matrix A is stable if all of its characteristic roots have negative real parts.*

Proceeding along these lines, one may be interested in how fast the system returns to equilibrium. To solve this problem we appeal to the characteristic roots of A. For example, suppose that A is diagonalizable with all characteristic roots negative, say

$$A = T^{-1} \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n) T. \tag{2}$$

Then, the solution of (1) is

$$e^{At} = T^{-1} \operatorname{diag}(e^{\lambda_1 t}, \cdots, e^{\lambda_n t}) T. \tag{3}$$

Now, if λ_k is the smallest characteristic root of A in absolute value, we see that the system returns to equilibrium at the rate $O(e^{\lambda_k t})$ as $t \to \infty$. Theorem A is proved for general matrices in [1] by first triangularizing the matrix; using this same approach, one can show that the system (1) defined by a general matrix whose characteristic roots have negative real part also returns to equilibrium at the rate $O(e^{\lambda_k t})$ as $t \to \infty$, where λ_k is the characteristic root which is smallest in absolute value.

Next, one may ask, "Does there exist a class of matrices such that each solution to (1) defined by this class converges "uniformly" to equilibrium?" By "uniformly", we mean that all the components of the solution approach equilibrium at a particular rate or faster. In answer to this question, let A be an M-matrix whose inverse is stochastic. It is well known that if A is an M-matrix, then $-A$ is stable. By Theorem 1 in § 3 below all the characteristic roots of $-A$ satisfy

$$\operatorname{Re} \lambda \leqq -1. \tag{4}$$

* Received by the editors December 18, 1978, and in revised form September 30, 1979.
† Department of Mathematics, University of Tennessee, Chattanooga, Tennessee 37402.

Hence the linear system

$$\frac{dx}{dt} = -Ax, \qquad x(0) = c \tag{5}$$

returns to equilibrium at the rate of $O(e^{-t})$, since -1 is the smallest characteristic root of $-A$ in absolute value. Thus we see that the class of M-matrices whose inverses are stochastic is an answer to our question.

This paper examines those M-matrices whose inverses are stochastic. Characterizations are given for such matrices, and properties of this class of matrices are determined.

2. Preliminaries. Throughout this paper, all matrices considered are $n \times n$ and real. A^T will denote the transpose of the matrix A, sp A will denote the spectrum of A, and det A will denote the determinant of A. We state the following definitions:

DEFINITION 1 (Ostrowski). A is an *M-matrix* if $a_{ij} \leqq 0$, $i \neq j$, and A possesses one of the following equivalent properties:

(a) A is nonsingular and the elements of A^{-1} are nonnegative.

(b) All principal minors of A are positive.

(c) There exist n positive numbers x_j such that

$$\sum_{j=1}^{n} a_{ij}x_j > 0, \qquad i = 1, \cdots, n.$$

DEFINITION 2. A is a *singular M-matrix* [3] if A is singular, $a_{ij} \leqq 0$ for $i \neq j$ and A has all principal minors nonnegative.

DEFINITION 3. Let A be an arbitrary $m \times n$ matrix. The *Moore–Penrose inverse* [2] *of* A is the unique $n \times m$ matrix A^+ satisfying $AA^+A = A$, $A^+AA^+ = A^+$, $(AA^+)^T = AA^+$ and $(A^+A)^T = A^+A$.

DEFINITION 4. A square matrix S is called *stochastic* [4] if S is nonnegative and if the sum of the elements of each row of S is 1.

DEFINITION 5. Let A be an $n \times n$ complex matrix and let sp $A = \{\lambda_1, \lambda_2, \cdots, \lambda_n\}$. The *spectral radius of* A, denoted $\rho(A)$, is defined by

$$\rho(A) = \max_{1 \leqq i \leqq n} |\lambda_i|.$$

DEFINITION 6. Let A and B be $n \times n$ real matrices. $A \geqq B$ means that $A-B$ is nonnegative.

3. Results. The first theorem characterizes those M-matrices whose inverses are stochastic.

THEOREM 1. *Suppose A is an M-matrix. Then the following statements are equivalent:*

(1) A^{-1} *is stochastic.*

(2) $Ae = e$ *where* $e = (1, 1, \cdots, 1)^T$.

(3) $A = \lambda_0 I - (\lambda_0 - 1)S$, *for some* $\lambda_0 > 1$ *and some stochastic matrix* S.

(4) Re $\lambda \geqq 1$ for all $\lambda \in$ sp A, *and* 1 *is a root of minimum modulus with corresponding eigenvector* e.

(5) $A = LU$ *where* L *is a lower triangular M-matrix with a stochastic inverse and* U *is an upper triangular M-matrix with a stochastic inverse.*

Proof. We will prove the implications as follows. First we will prove $(i) \Leftrightarrow (j)$ for i, $j \leqq 4$. Then we will show (2) $\Leftrightarrow$ (5).

(1)$\Rightarrow$(2). If A^{-1} is stochastic, then $A^{-1}e = e$. Hence $Ae = A(A^{-1}e) = (AA^{-1})e = e$.

(2)$\Rightarrow$(3). It is well known that if A is an M-matrix, $A = \lambda_0 I - B$ for some nonnegative matrix B and some $\lambda_0 > \rho(B)$. Note that we may choose $\lambda_0 > 1$. Hence

$$Be = (\lambda_0 I - A)e = \lambda_0 e - e = (\lambda_0 - 1)e.$$

Thus $S = (\lambda_0 - 1)^{-1}B$ is stochastic and $A = \lambda_0 I - (\lambda_0 - 1)S$.

(3)$\Rightarrow$(4). Suppose that $A = \lambda_0 I - (\lambda_0 - 1)S$ for some $\lambda_0 > 1$ and some stochastic matrix S. Then, if $\operatorname{sp} S = \{\lambda_1, \lambda_2, \cdots, \lambda_n\}$,

$$\operatorname{sp} A = \{\lambda_0 - (\lambda_0 - 1)\lambda_1, \lambda_0 - (\lambda_0 - 1)\lambda_2, \cdots, \lambda_0 - (\lambda_0 - 1)\lambda_n\}.$$

It is well known that if $\lambda \in \operatorname{sp} S$, then $|\lambda| \leqq 1$. This implies that $\operatorname{Re} \lambda \leqq 1$ for all $\lambda \in \operatorname{sp} S$. Thus $\operatorname{Re}[\lambda_0 - (\lambda_0 - 1)\lambda_i] \geqq \operatorname{Re}[\lambda_0 - (\lambda_0 - 1)] \geqq 1$, $1 \leqq i \leqq n$. Therefore, if $\lambda \in \operatorname{sp} A$, $\operatorname{Re} \lambda \geqq 1$. Further, 1 is a root of A with corresponding eigenvector e, since $Ae = [\lambda_0 1 - (\lambda_0 - 1)S]e = e$. Hence 1 is a root of minimum modulus.

(4)$\Rightarrow$(1). Since 1 is a root of A with corresponding eigenvector e, $Ae = e$. Hence $e = A^{-1}e$ so that A^{-1} is stochastic.

(5)$\Rightarrow$(2). Suppose $A = LU$, where L is a lower triangular M-matrix with a stochastic inverse and U is an upper triangular M-matrix with a stochastic inverse. Then $Ae = LUe = Le = e$.

(2)$\Rightarrow$(5). Suppose $Ae = e$. Fiedler and Ptak [3, Thm. 4.3] have shown that each M-matrix A has a factorization $A = L_1 U_1$, where $L_1 = (l_{ij})$ is a lower triangular M-matrix and $U_1 = (u_{ij})$ is an upper triangular M-matrix. Let $D = \operatorname{diag}(d_1, d_2, \cdots, d_n)$, where $d_i = \sum_{j=1}^{n} u_{ij} > 0$, $1 \leqq i \leqq n$, and let $U = D^{-1}U_1$. Then, $Ue = e$ and U is an upper triangular M-matrix with a stochastic inverse. Further, $e = Ae = L_1 DUe = L_1 De$, which implies that $L = L_1 D$ is a lower triangular M-matrix with a stochastic inverse.

Remark. Kuo [6, Thm. 3.9] has shown that if A is a singular M-matrix, then there exists a permutation matrix P such that

$$PAP^T = \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix},$$

where A is an M-matrix. Hence

$$PA^+P^T = \begin{bmatrix} A_1^{-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

Thus, we see that a singular M-matrix cannot have a stochastic Moore–Penrose inverse.

If A is an M-matrix, $\det A > 0$. In the following corollary, we are able to make a stronger statement provided that A has a stochastic inverse.

COROLLARY 1. *If A is an M-matrix with a stochastic inverse, then* $\det A \geqq 1$.

Proof. Since A^{-1} is stochastic, the spectrum of A^{-1} lies in the unit circle, which in turn implies that $|\det A^{-1}| = |\prod_{i=1}^{n} \lambda_i| \leqq 1$. Hence $|\det A| \geqq 1$. Also, $\det A > 0$, since A is an M-matrix. Thus, $\det A \geqq 1$.

Remark. It is to be noted that we only needed the hypotheses that A^{-1} is stochastic and $\det A > 0$. Also, the condition that $\det A \geqq 1$ is not sufficient; for example, the matrix $A = 2I$ is an M-matrix such that $\det A \geqq 1$, but obviously $A^{-1} = \frac{1}{2}I$ is not stochastic.

Let $A = (a_{kj})$ be a square matrix of order n with real or complex elements. Let

$$P_k = \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{kj}|, \qquad k = 1, 2, \cdots, n. \tag{6}$$

It is well known that each characteristic root of A lies in the interior or on the boundary of at least one of the circles

$$C_k: \quad |z - a_{kk}| \leqq P_k, \qquad k = 1, 2, \cdots, n. \tag{7}$$

We apply this result to prove the following corollary.

COROLLARY 2. *Suppose A is an M-matrix whose inverse is stochastic. Then the spectrum of A is contained in the circular region with center a and radius $a-1$, where $a = \max_i a_{ii}$.*

Proof. Note that $P_k = a_{kk} - 1$ for $k = 1, 2, \cdots, n$. Thus (7) reduces to

$$C_k: \quad |z - a_{kk}| \leqq a_{kk} - 1, \qquad k = 1, 2, \cdots, n. \tag{8}$$

Now (8) is a set of circles satisfying $C_k \subseteq C_j$ or $C_j \subseteq C_k$ for $1 \leqq j, k \leqq n$, and each circle contains 1 as a boundary point. The largest of these circles is

$$C_l: \quad |z - a_{ll}| \leqq a_{ll} - 1, \quad \text{where } a_{ll} = a = \max_i a_{ii}.$$

Thus the spectrum of A is contained in the circular region with center a and radius $a-1$.

It is well known that the principal minors of an M-matrix are positive and that the real part of each eigenvalue of an M-matrix is positive. We now obtain sharper results for M-matrices whose inverses are stochastic.

LEMMA 1. *Suppose that A is an M-matrix whose inverse is stochastic and $B = (b_{ij})$ is a matrix with nonpositive offdiagonal elements. Then, if $B \geqq A$, $\operatorname{Re} \lambda \geqq 1$ for all $\lambda \in \operatorname{sp} B$.*

Proof. Let $\lambda \in \operatorname{sp} B$, say $Bx = \lambda x$, where $x \neq 0$. Further, let $M = |x_k| = \max_{1 \leqq i \leqq n} |x_i| > 0$. Now $(B - \lambda I)x = 0$ implies that $(b_{kk} - \lambda)x_k = \sum' (-b_{kj}x_j)$, where the prime means we sum over $j \neq k$. Hence

$$\begin{aligned} M[\operatorname{Re}(b_{kk} - \lambda)] \leqq M|b_{kk} - \lambda| &= |x_k||b_{kk} - \lambda| \\ &\leqq \sum' |b_{kj}||x_j| \\ &= \sum' (-b_{kj})|x_j| \\ &\leqq M \sum' (-b_{kj}). \end{aligned}$$

Thus $b_{kk} - \operatorname{Re} \lambda \leqq \sum' (-b_{kj})$. Since A has a stochastic inverse and $B \geqq A$, $Be \geqq Ae = e$. Thus

$$1 \leqq \sum_{j=1}^{n} b_{kj} \leqq \operatorname{Re} \lambda.$$

THEOREM 2. *Let $A = (a_{ij})$ be an M-matrix with a stochastic inverse. Then, if A' is a principal submatrix of A,*

i) $\operatorname{Re} \lambda \geqq 1$ *for all* $\lambda \in \operatorname{sp} A'$ *and*

ii) $\det A' \geqq 1$.

Proof. i) Let A' be a principal $k \times k$ submatrix of A. Without loss of generality, we may assume that

$$A = \begin{bmatrix} A' & B \\ C & D \end{bmatrix}.$$

Now let

$$B' = \begin{bmatrix} A' & & 0 \\ & a_{k+1,k+1} & \\ & & \ddots \\ 0 & & a_{n,n} \end{bmatrix} \geqq A.$$

By Lemma 1, $\operatorname{Re} \lambda \geqq 1$ for all $\lambda \in \operatorname{sp} B'$ which implies $\operatorname{Re} \lambda \geqq 1$ for all $\lambda \in \operatorname{sp} A'$.

ii) If λ_i is real, $\lambda_i \geqq 1$ by i). If λ_i is imaginary, $\lambda_i = \bar{\lambda}_j$ for some $i \neq j$, since A' is a real matrix. This implies that $\lambda_i \bar{\lambda}_j \geqq 1$ by i). Hence, $\det A' = \prod_{i=1}^{k} \lambda_i \geqq 1$.

Remark. In Corollary 1 we noted that the conclusion followed without the assumption that A is an M-matrix. In Theorem 2 this is not the case. For example, consider

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

Note that $\det A \geqq 1$, and

$$A^{-1} = \frac{1}{6} \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 4 \\ 2 & 2 & 2 \end{bmatrix}$$

is stochastic. However, it is obvious that $\det A' \ngeqq 1$ for all principal submatrices A' of A. Further, consider

$$B = \begin{bmatrix} 12 & -2 & -9 \\ -12 & 4 & 9 \\ 6 & -2 & -3 \end{bmatrix}.$$

Then

$$B^{-1} = \frac{1}{6} \begin{bmatrix} 1 & 2 & 3 \\ 3 & 3 & 0 \\ 0 & 2 & 4 \end{bmatrix}$$

is stochastic and the roots of B are 1 and the double root 6. On the other hand, it is easily seen that $\operatorname{Re} \lambda \ngeqq 1$ for all $\lambda \in \operatorname{sp} B'$ where B' is a principal submatrix of B.

Next we will show that if an M-matrix A has a stochastic inverse, then $A + A^T$ is an M-matrix.

THEOREM 3. *Let A be a positive diagonally dominant M-matrix. Then, $A + A^T$ is a positive definite M-matrix.*

Proof. We shall prove the theorem by induction on the size n of A. If $n = 1$, then obviously the theorem is true. So assume the theorem is true for all positive diagonally dominant M-matrices of size r, where $1 \leqq r < k$. Let A be a $k \times k$ positive diagonally dominant M-matrix, $E = A + A^T$, and E' be a principal submatrix of E. If $E' = E$, then $\det E' = \det E > 0$, since it is well known that a positive diagonally dominant matrix has a positive determinant. So we assume that E' is a proper submatrix of E, and without loss of generality we may assume that

$$E = \begin{bmatrix} E' & E_2 \\ E_3 & E_4 \end{bmatrix} = \begin{bmatrix} A' + (A')^T & B + C^T \\ C + B^T & D + D^T \end{bmatrix}$$

where $A = \begin{bmatrix} A' & B \\ C & D \end{bmatrix}$. A' is a positive diagonally dominant M-matrix since A is, and $A' + (A')^T$ is positive definite by the induction hypothesis. Hence, $\det E' = \det (A' + (A')^T) > 0$. Thus, $A + A^T$ is an M-matrix, since it obviously has the required sign pattern. Immediately we have

COROLLARY 3. *If A is an M-matrix whose inverse is stochastic, then $A + A^T$ is a positive definite M-matrix.*

Our final results pertain to factorizations with M-matrices whose inverses are stochastic.

THEOREM 4. *The M-matrix A is positive diagonally dominant if and only if $A = DS$ where D is a positive diagonal matrix and S is an M-matrix with a stochastic inverse.*

Proof. First, assume that $A = DS$, where D is a positive diagonal matrix and S is an M-matrix with a stochastic inverse. By Theorem 1 and Definition 1, $Ae = DSe = De > 0$, which implies that A is a positive diagonally dominant M-matrix. Conversely, suppose A is a positive diagonally dominant M-matrix. Then, $a_{ii} > \sum_{j=1, j\neq i}^{n} (-a_{ij})$. Let $\sum_{j=1}^{n} a_{ij} = d_i > 0$, $1 \leqq i \leqq n$. Then

$$\sum_{j=1}^{n} \frac{a_{ij}}{d_i} = 1, \qquad 1 \leqq i \leqq n.$$

So $A = DS$, where $D = \operatorname{diag}(d_1, d_2, \cdots, d_n)$, and $S = (s_{ij})$, where $s_{ij} = a_{ij}/d_i$, $1 \leqq i, j \leqq n$. Note that $Se = e$, which implies that S is an M-matrix with a stochastic inverse.

COROLLARY 4. *Let A be an M-matrix. Then there exist positive diagonal matrices D_1 and D_2 and an M-matrix S with a stochastic inverse such that $A = D_1 S D_2$.*

Proof. It is well known that there exists a positive diagonal matrix D_2 such that $A = WD_2$, where W is an M-matrix with dominant positive principle diagonal. The corollary follows with Theorem 4 applied to W.

Note that the above corollary implies that the inverse of each M-matrix is diagonally equivalent to a stochastic matrix.

THEOREM 5. *Suppose that A is an M-matrix with $q(A) > 0$ a root of minimum modulus. Then, there exists a positive diagonal matrix D such that $A = Dq(A)SD^{-1}$, where S is an M-matrix with a stochastic inverse provided that A is irreducible.*

Proof. Suppose that A is an M-matrix with $q(A) > 0$ a root of minimum modulus. Then $A^{-1} = (a'_{ij})$ is a nonnegative matrix with $1/q(A)$ a root of maximum modulus. By the Perron–Frobenius theorem, there exists $x > 0$ such that $A^{-1}x = (1/q(A))x$ since A is irreducible. If $x = (x_1, x_2, \cdots, x_n)$, this implies that

$$\frac{1}{q(A)} x_i = \sum_{j=1}^{n} a'_{ij} x_j. \tag{*}$$

Let $D = \operatorname{diag}(x_1, x_2, \cdots, x_n)$ and let $S_1 = q(A)D^{-1}A^{-1}D = (s_{ij})$. Then $s_{ij} = q(A)x_i^{-1}a'_{ij}x_j \geqq 0$, and

$$\sum_{j=1}^{n} s_{ij} = q(A)x_i^{-1} \sum_{j=1}^{n} a'_{ij}x_j = 1$$

by (*). Thus S_1 is stochastic, which implies that $S_1^{-1} = (1/q(A))D^{-1}AD$ is an M-matrix with a stochastic inverse, since $S_1^{-1}D^{-1}x > 0$ and S_1^{-1} has the required sign pattern. Therefore, $A = Dq(A)S_1^{-1}D^{-1}$ and the theorem holds.

THEOREM 6. *Suppose that A is a symmetric M-matrix whose inverse is stochastic. Then, there exists an upper triangular M-matrix S with a stochastic inverse and a diagonal matrix D with positive diagonal entries such that $A = S^T DS$. Further, $S^T D$ is a lower triangular M-matrix with a stochastic inverse.*

Proof. Jacobson [5, Thm. 1] has shown that each symmetric M-matrix A has a factorization $A = GG^T$, where G is a lower triangular M-matrix. Let $G^T = (g'_{ij}) = D_1 S$, where $D_1 = \text{diag}\,(d_1, d_2, \cdots, d_n)$ and $d_i = \sum_{j=1}^{n} g'_{ij} > 0$, $1 \leqq i \leqq n$. Then S is an M-matrix with a stochastic inverse by (2) of Theorem 1. Therefore $A = GG^T = (S^T D_1)(D_1 S) = S^T D S$, where $D = D_1^2$. It is easily shown that $S^T D$ is an M-matrix with a stochastic inverse.

REFERENCES

[1] R. BELLMAN, *Introduction to Matrix Analysis*, McGraw-Hill, New York. 1969.

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

[3] M. FIEDLER AND V. PTÁK, *On matrices with non-positive off-diagonal elements and positive principal minors*, Czechoslovak Math. J., 12 (1962), pp. 382–400.

[4] F. R. GANTMACHER, *The Theory of Matrices*, vol. 2, Chelsea, New York, 1959.

[5] D. H. JACOBSON, *Factorization of symmetric M-matrices*, Linear Algebra Appl., 9 (1974), pp. 275–278.

[6] I. KUO. *The Moore–Penrose inverse of singular M-matrices*, Linear Algebra and Appl., 17 (1977), pp. 1–14.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 3, September 1981

0196-5212/81/0203-0006 $01.00/0

THE $v \times v$ $(0, 1, -1)$-CIRCULANT EQUATION $AA^T = vI - J$*

JAMES H. McKAY† AND STUART SUI-SHENG WANG†

Abstract. An electromechanical pulse generator has been proposed (J. P. Craig and R. Saeks, *An electromechanical pulse generator*, Proc. 1st IEEE International Pulsed Power Conference, Institute of Electrical and Electronics Engineers, 1976, pp. IIB 7-1–IIB 7-4) which is equivalent to finding a $v \times v$ circulant matrix A with entries from $\{0, 1, -1\}$ such that $AA^T = vI - J$. In this earlier work it is reported that if v is an odd prime and the entries in the first row of A are the Legendre symbols (j/v), $0 \leqq j \leqq v-1$, then A is a solution. It was conjectured that A exists only if v is an odd prime and that the solution is unique up to cyclic permutation of the columns and multiplication of A by -1. In this paper, using convolution products, Fourier transforms and number theory, we settle these two conjectures affirmatively.

1. Introduction. The usual techniques for generating high power electrical pulses of short pulse duration and high recurrence frequency depend upon the storage of electrical energy either in an electrostatic field or in a magnetostatic field, and the subsequent discharge of a fraction or all of this stored energy into the load [8]. Another approach is to employ an electromechanical energy converter to convert mechanically stored energy into the desired electrical pulses. A report concerning a feasibility study of some possible schemes and problems associated with them was made in [6]. A new scheme was proposed by Craig and Saeks [7], which gives a lower pulse repetition rate than a conventional electromechanical pulse generator (where the north and south poles are in alternating positions) with a comparable number of poles. Such a generator is designed with a prime number of poles, p. The poles are numbered 0 through $p-1$. The zeroth pole carries zero flux, and the remaining poles are divided equally between north and south poles, each carrying an equal amount of flux. The positions of the $(p-1)/2$ north poles are determined from a Legendre sequence of modulo p. The north poles correspond to the perfect squares of modulo p, and the remaining nonzero poles are the south poles. It was proven that such a generator will produce one positive pulse per revolution that is $(p-1)$ times as large as $(p-1)$ negative pulses which are produced [4]. The actual design and construction of an eleven-pole pulser were carried out and its output was measured [3].

However, it was not known whether (1) only prime p will work; and (2) the arrangements of north and south poles will be unique.

If a north pole is symbolized by "+1", a south pole by "−1", and a neutral one by "0", then the desired physical properties can be described mathematically as follows. The purpose is to get a sequence $[a_0, a_1, \cdots, a_{v-1}]$ of v terms, each a_j either $+1$ or -1 or 0, such that

$$\begin{aligned} a_0^2 + a_1^2 + a_2^2 + \cdots + a_{v-1}^2 &= v - 1, \\ a_{v-1}a_0 + a_0a_1 + a_1a_2 + \cdots + a_{v-2}a_{v-1} &= -1, \\ a_{v-2}a_0 + a_{v-1}a_1 + a_0a_2 + \cdots + a_{v-3}a_{v-1} &= -1, \\ &\vdots \\ a_1a_0 + a_2a_1 + a_3a_2 + \cdots + a_0a_{v-1} &= -1. \end{aligned} \tag{1.1}$$

Henceforth, the problem of finding a sequence satisfying the properties above will be referred to as the *binary coded pulser problem* of order v. It was known that if v is an odd

* Received by the editors May 5, 1980, and in final form January 29, 1981.
† Department of Mathematical Sciences, Oakland University, Rochester, Michigan 48063.

prime, then the sequence formed by the Legendre symbols is a solution to the binary coded pulser problem [7]. Saeks made the following two conjectures:

CONJECTURE 1. *If the binary coded pulser problem of order v has a solution, then v is a prime.*

CONJECTURE 2. *If the binary coded pulser problem of order v has a solution, then the solution is essentially unique.*

These two conjectures have been verified up to $v=41$ by computer [4]. The purpose of this paper is to settle these two conjectures affirmatively.

2. Preliminaries. The matrices in this paper are either $v\times v$ or $1\times v$ over $\mathbb{C}$, the field of complex numbers, and their entries are indexed by residues modulo v, $0, 1, 2, \cdots, v-1$. For any sequence $\mathbf{a}=[a_0, a_1, \cdots, a_{v-1}]$, $\mathbf{a}^{-}=[\bar{a_0}, \bar{a_1}, \cdots, \bar{a_{v-1}}]$ denotes the complex conjugate of $\mathbf{a}$, $\mathbf{a}^{\sim}=[a_0, a_{v-1}, a_{v-2}, \cdots, a_1]$, $P(\mathbf{a}, X)$ is the polynomial $a_0+a_1X+\cdots+a_{v-1}X^{v-1}$, $D(\mathbf{a})$ is the diagonal matrix diag $(a_0, a_1, \cdots, a_{v-1})$, and $M(\mathbf{a})$ is the *circulant* matrix

$$\begin{bmatrix} a_0 & a_1 & a_2 & \cdots & a_{v-1} \\ a_{v-1} & a_0 & a_1 & \cdots & a_{v-2} \\ a_{v-2} & a_{v-1} & a_0 & \cdots & a_{v-3} \\ & & \cdots & & \\ a_1 & a_2 & a_3 & \cdots & a_0 \end{bmatrix}.$$

Evidently $M(\mathbf{a}^{-})=M(\mathbf{a})^{-}$ and $M(\mathbf{a}^{\sim})=M(\mathbf{a})^{T}$, where $M(\mathbf{a})^{T}$ is the transpose of $M(\mathbf{a})$. Furthermore, if we let C denote the permutation (also circulant) matrix with 1's in positions $(0, 1), (1, 2), \cdots, (v-2, v-1), (v-1, 0)$ and 0's elsewhere, then $M(\mathbf{a})=P(\mathbf{a}, C)$.

It can be easily verified that the system of equations (1.1) in the binary coded pulser problem is equivalent to a single circulant matrix equation

$$M(\mathbf{a})M(\mathbf{a})^{T}=vI-J, \tag{2.1}$$

where I is the identity matrix and J is the matrix with 1's in all positions. Equation (2.1) can be written in a more concise way once we introduce the concept of convolution between sequences.

The *convolution* of $\mathbf{a}$ and $\mathbf{b}$, denoted by $\mathbf{a}*\mathbf{b}$, is defined by the following formula. If $\mathbf{c}=\mathbf{a}*\mathbf{b}$, then

$$c_\gamma=\sum_{\alpha+\beta=\gamma} a_\alpha b_\beta, \qquad \gamma=0, 1, 2, \cdots, v-1.$$

Then sequences together with term-by-term addition and convolution form a commutative ring with identity. A moment's reflection will convince one that this commutative ring with identity is isomorphic to each one of the following:

(i) the subring of the matrix ring, consisting of circulant matrices;

(ii) the quotient ring $\mathbb{C}[x]=\mathbb{C}[X]/\langle X^v-1\rangle$, where X is an indeterminate over $\mathbb{C}$, $\langle X^v-1\rangle$ is the principal ideal generated by X^v-1 and x is the coset of X;

(iii) the group ring $\mathbb{C}[G]$ of a cyclic group $G=\{1, g, g^2, \cdots, g^{v-1}\}$ of order v.

(The correspondence is $\mathbf{a}\leftrightarrow P(\mathbf{a}, C)=M(\mathbf{a})\leftrightarrow P(\mathbf{a}, x)\leftrightarrow P(\mathbf{a}, g)$.)

On the other hand, sequences together with term-by-term addition and term-by-term multiplication (denoted by "$\circ$") form a commutative ring with identity which is isomorphic to each one of the following:

(i)′ the subring of the matrix ring consisting of diagonal matrices;

(ii)′ the direct product of v copies of $\mathbb{C}$.

(The correspondence is $\mathbf{a} \leftrightarrow D(\mathbf{a}) \leftrightarrow (a_0, a_1, \cdots, a_{v-1})$.)

Then it follows from the isomorphisms among (i), (ii) and (iii) that (2.1) is equivalent to

$$P(\mathbf{a}, x)P(\mathbf{a}, x^{-1}) = v - (1 + x + x^2 + \cdots + x^{v-1}), \tag{2.2}$$

which is also equivalent to

$$P(\mathbf{a}, g)P(\mathbf{a}, g^{-1}) = v - (1 + g + g^2 + \cdots + g^{v-1}). \tag{2.3}$$

In other words, one can use either (2.1) or (2.2) or (2.3) to study the binary coded pulser problem. In the past, these three approaches have been taken to attack other problems: Hall and Ryser [11], [12] used polynomials in the indeterminates X, X^{-1} and double modulus arguments; Bruck [2] used elements in the group ring of a cyclic group; Newman [16] used circulant matrices to study multipliers of cyclic difference sets.

When viewing the solution as an element in the ring $\mathbb{Q}[X]/\langle X^v - 1\rangle$, it is natural to use the Chinese remainder theorem to decompose the ring into a direct product of fields,

$$\frac{\mathbb{Q}[X]}{\langle X^v - 1\rangle} \cong \prod_{d|v} \frac{\mathbb{Q}[X]}{\langle \Phi_d(X)\rangle} \cong \prod_{d|v} \mathbb{Q}[\omega^{v/d}], \tag{2.4}$$

where $\mathbb{Q}$ denotes the field of rational numbers, $\Phi_d(X)$ denotes the dth cyclotomic polynomial and $\omega = e^{2\pi i/v}$. Ring theoretically, we can recover the global information, i.e., an element in $\mathbb{Q}[X]/\langle X^v - 1\rangle$, from the local information in each of the fields $\mathbb{Q}[\omega^{v/d}]$. When viewing the solution as a sequence, it is natural to utilize the properties of the Fourier transform because of the cyclic nature of the solution. The process of recovering from the local information to the global information is easier in the Fourier transform approach than in the ring theoretic approach. For example, in the case $v = 6$, an element $(f(1), f(\omega^3), f(\omega^2), f(\omega))$ of $\mathbb{Q}[1] \times \mathbb{Q}[\omega^3] \times \mathbb{Q}[\omega^2] \times \mathbb{Q}[\omega]$ corresponds to an element of $\mathbb{Q}[X]/\langle X^6 - 1\rangle$ by the Chinese remainder theorem. However, $f(\omega^5)$ and $f(\omega^4)$ are easily found from $f(\omega)$ and $f(\omega^2)$, respectively, by appropriate automorphisms and then the element $f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5$ in $\mathbb{Q}[X]/\langle X^6 - 1\rangle$ is recovered by applying the Fourier transform to $(1/\sqrt{6})[f(1), f(\omega), f(\omega^2), f(\omega^3), f(\omega^4), f(\omega^5)]$. Consequently, the principal approach used in this paper is sequences, convolution and Fourier transform.

The result of this paper can provide a simple alternative proof of theorems of Kelly [13], which extended results of Perron [17].

3. Binary coded pulser problem. Equations (1.1) or (2.1) can be reformulated in terms of the convolution product:

DEFINITION 3.1. A *binary coded pulser of order* v is a sequence $\mathbf{a} = [a_0, a_1, a_2, \cdots, a_{v-1}]$ of v terms composed entirely of 0's, +1's and −1's such that $\mathbf{a} * \mathbf{a}^{\sim} = [v - 1, -1, -1, \cdots, -1]$.

Our first observation is

LEMMA 3.2. (i) *If* $\mathbf{a}$ *is a binary coded pulser of order* v, *then* v *is odd, exactly one of the* a_j *is zero and the other* a_j*'s are equally divided between* +1 *and* −1.

(ii) *If, furthermore* $a_0 = 0$, *then* $\mathbf{a}^{\sim} = (-1)^{(v-1)/2}\mathbf{a}$; *i.e.,* $a_{v-j} = (-1)^{(v-1)/2} a_j$ *for all* j.

Proof. (i) The zeroth term of $\mathbf{a} * \mathbf{a}^{\sim}$ is $\sum_{j=0}^{v-1} a_j^2 = v - 1$, which shows that exactly one of the a_j is zero. If s denotes the sum of the terms in $\mathbf{a}$, then the sum of the terms in $\mathbf{a} * \mathbf{a}^{\sim}$ is $s^2 = (v - 1) - 1 - 1 - \cdots - 1 = 0$ and $s = 0$.

(ii) The jth term of $\mathbf{a} * \mathbf{a}^{\sim}$, with $j \neq 0$, is $\sum_{\alpha=0}^{v-1} a_\alpha a_{j+\alpha} = -1$.

Two of $a_0a_j, a_1a_{j+1}, a_2a_{j+2}, \cdots, a_{v-1}a_{j+v-1}$ are equal to zero. They are a_0a_j and $a_{v-j}a_0$. The remaining are ± 1 and, since they total -1, there must be exactly $(v-1)/2$ of them equal to -1 and the product of these is $(-1)^{(v-1)/2}$. The proof is completed by noting that $1 = \prod_{\alpha=1}^{v-1} a_\alpha^2 = a_j a_{v-j}(-1)^{(v-1)/2}$. □

DEFINITION 3.3. A binary coded pulser $\mathbf{a} = [a_0, a_1, a_2, \cdots, a_{v-1}]$ is called a *special* (resp. *normalized*) binary coded pulser if $a_0 = 0$ (resp. $a_0 = 0$ and $a_1 = 1$).

Thus a special binary coded pulser of order v is a sequence $\mathbf{a} = [a_0, a_1, a_2, \cdots, a_{v-1}]$ such that $\mathbf{a} \circ \mathbf{a} = [0, 1, 1, \cdots, 1]$ and $\mathbf{a} * \mathbf{a}^{\sim} = [v-1, -1, -1, \cdots, -1]$.

Clearly, if there is a binary coded pulser $\mathbf{a}$ of order v, then there is a normalized binary coded pulser of order v which is formed from $\mathbf{a}$ by a cyclic permutation and multiplication by -1, if necessary.

A sequence $\mathbf{a} = [a_0, a_1, a_2, \cdots, a_{v-1}]$ is called *even* if $\mathbf{a} = \mathbf{a}^{\sim}$, and is called *odd* if $\mathbf{a} = -\mathbf{a}^{\sim}$. Note that $\mathbf{a}$ is even if and only if the corresponding circulant matrix $M(\mathbf{a})$ is symmetric; $\mathbf{a}$ is odd if and only if $M(\mathbf{a})$ is skew-symmetric. Thus, Lemma 3.2 (ii) says that if $\mathbf{a}$ is a special (in particular, normalized) binary coded pulser of order v, then $M(\mathbf{a})$ is either symmetric or skew-symmetric depending on $v \equiv 1$ or 3 (mod 4), respectively.

Recall that the *Jacobi symbol* (a/b) is defined for all odd a, b by the following: (i) if $(a, b) = 1$ and $b = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_r^{\alpha_r}$ is the prime factorization of b, then $(a/b) = \prod_{j=1}^{r} (a/p_j)^{\alpha_j}$ where (a/p_j) is the Legendre symbol; (ii) if $(a, b) > 1$, then $(a/b) = 0$. It is then easily verified that the Jacobi symbol is *bimultiplicative* in the sense that

$$\left(\frac{a}{b_1}\right)\left(\frac{a}{b_2}\right) = \left(\frac{a}{b_1 b_2}\right) \quad \text{and} \quad \left(\frac{a_1}{b}\right)\left(\frac{a_2}{b}\right) = \left(\frac{a_1 a_2}{b}\right);$$

and furthermore $(-1/b) = (-1)^{(b-1)/2}$ [10, pp. 76–77].

There are a few special matrices which we use:

$$W = (\omega^{\alpha\beta}), \quad \alpha = 0, 1, \cdots, v-1, \quad \beta = 0, 1, \cdots, v-1, \quad \text{where } \omega = e^{2\pi i/v},$$

$$D = \operatorname{diag}(1, \omega, \omega^2, \cdots, \omega^{v-1}),$$

$$E = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \cdots & & 1 \\ \vdots & & \iddots & \vdots \\ 0 & 1 & \cdots & 0 \end{bmatrix}.$$

Note that W is the character table of a cyclic group of order v. Also $\mathbf{a}E = \mathbf{a}^{\sim}$, so that right multiplication of sequences by E corresponds to the involution of the group ring $\mathbb{C}[G]$ induced by $g \mapsto g^{-1}$. It follows from the orthogonality relations among characters that

$$WW^{-} = W^{-}W = vI \tag{3.1}$$

(which can also be easily verified using the summation formula for geometric series). It is also easily verified that $W^2 = vE$, $W = EW^{-} = W^{-}E$, $ECE = C^T = C^{v-1} = C^{-1}$, $ED(\mathbf{a})E = D(\mathbf{a}^{\sim})$, $CW = WD$ and $WC = D^{-}W$. As a consequence of $W^{-1}CW = D$ and $M(\mathbf{a}) = P(\mathbf{a}, C)$,

$$W^{-1}M(\mathbf{a})W = \operatorname{diag}(P(\mathbf{a}, 1), P(\mathbf{a}, \omega), P(\mathbf{a}, \omega^2), \cdots, P(\mathbf{a}, \omega^{v-1})), \tag{3.2}$$

which gives the eigenvalues of a circulant matrix and the classical result that $\det(M(\mathbf{a})) = P(\mathbf{a}, 1)P(\mathbf{a}, \omega)P(\mathbf{a}, \omega^2)\cdots P(\mathbf{a}, \omega^{v-1})$. Similarly, we have

$$WM(\mathbf{a})W^{-1} = \operatorname{diag}(P(\mathbf{a}, 1), P(\mathbf{a}, \omega^{-1}), P(\mathbf{a}, \omega^{-2}), \cdots, P(\mathbf{a}, \omega^{-(v-1)})). \tag{3.3}$$

We define the *Fourier transform* $\mathbf{a}^\wedge$ of $\mathbf{a}$ by

$$\mathbf{a}^\wedge = \sqrt{v}\,\mathbf{a}W^{-1} = \frac{1}{\sqrt{v}}[P(\mathbf{a}, 1), P(\mathbf{a}, \omega^{-1}), P(\mathbf{a}, \omega^{-2}), \cdots, P(\mathbf{a}, \omega^{-(v-1)})],$$

and the *inverse Fourier transform* $\mathbf{a}^\vee$ of $\mathbf{a}$ by

$$\mathbf{a}^\vee = \frac{1}{\sqrt{v}}\,\mathbf{a}W = \frac{1}{\sqrt{v}}[P(\mathbf{a}, 1), P(\mathbf{a}, \omega), P(\mathbf{a}, \omega^2), \cdots, P(\mathbf{a}, \omega^{v-1})].$$

(Note that our definition of Fourier transform is slightly different from the one defined in [5], where $\mathbf{a}W^{-1}$ is called the Fourier transform of $\mathbf{a}$.) It is immediate that $\mathbf{a}^{\wedge\vee} = \mathbf{a} = \mathbf{a}^{\vee\wedge}$, $\mathbf{a}^{\wedge\wedge} = \mathbf{a}^{\sim} = \mathbf{a}^{\vee\vee}$, $\mathbf{a}^\wedge = \mathbf{a}^{\sim\vee} = \mathbf{a}^{\vee\sim}$, $\mathbf{a}^\vee = \mathbf{a}^{\sim\wedge} = \mathbf{a}^{\wedge\sim}$, $\mathbf{a}^{\wedge\wedge\wedge\wedge} = \mathbf{a} = \mathbf{a}^{\vee\vee\vee\vee}$, $\mathbf{a}^{-\wedge} = \mathbf{a}^{\vee-}$ and $\mathbf{a}^{-\vee} = \mathbf{a}^{\wedge-}$. The terms of $\sqrt{v}\,\mathbf{a}^\wedge$ (resp. $\sqrt{v}\,\mathbf{a}^\vee$) are precisely the eigenvalues of $M(\mathbf{a})$. Equations (3.2) and (3.3) can be rewritten as

$$WM(\mathbf{a})W^{-1} = \sqrt{v}D(\mathbf{a}^\wedge), \tag{3.4}$$

$$W^{-1}M(\mathbf{a})W = \sqrt{v}D(\mathbf{a}^\vee). \tag{3.5}$$

Therefore, taking $M(\mathbf{a}*\mathbf{b}) = M(\mathbf{a})M(\mathbf{b})$ and $D(\mathbf{a}\circ\mathbf{b}) = D(\mathbf{a})D(\mathbf{b})$ into account gives that

$$(\mathbf{a}*\mathbf{b})^\wedge = \sqrt{v}\,\mathbf{a}^\wedge \circ \mathbf{b}^\wedge, \tag{3.6}$$

$$(\mathbf{a}*\mathbf{b})^\vee = \sqrt{v}\,\mathbf{a}^\vee \circ \mathbf{b}^\vee, \tag{3.7}$$

$$(\mathbf{a}\circ\mathbf{b})^\wedge = \frac{1}{\sqrt{v}}\,\mathbf{a}^\wedge * \mathbf{b}^\wedge, \tag{3.8}$$

$$(\mathbf{a}\circ\mathbf{b})^\vee = \frac{1}{\sqrt{v}}\,\mathbf{a}^\vee * \mathbf{b}^\vee. \tag{3.9}$$

LEMMA 3.4. *Let* $\mathbf{a} = [a_0, a_1, a_2, \cdots, a_{v-1}]$ *be a special binary coded pulser and let* $(-1/v) = (-1)^{(v-1)/2}$ *be the Jacobi symbol. Then*

(i) $$P(\mathbf{a}, \omega^t)^2 = \begin{cases} 0 & \text{if } t \equiv 0 \pmod{v}, \\ \left(\dfrac{-1}{v}\right) v & \text{otherwise.} \end{cases}$$

(ii) $\sqrt{\left(\dfrac{-1}{v}\right)}\,\mathbf{a}^\vee$ *and* $\sqrt{\left(\dfrac{-1}{v}\right)}\,\mathbf{a}^\wedge$ *are special binary coded pulsers.*

Proof. (i) From Lemma 3.2 (ii) and (3.7) it follows that

$$\mathbf{a}^\vee \circ \mathbf{a}^\vee = \left(\frac{-1}{v}\right)\mathbf{a}^\vee \circ \mathbf{a}^{\sim\vee} = \left(\frac{-1}{v}\right)\frac{1}{\sqrt{v}}(\mathbf{a}*\mathbf{a}^\sim)^\vee = \left(\frac{-1}{v}\right)\frac{1}{\sqrt{v}}[v-1, -1, -1, \cdots, -1]^\vee$$

$$= \left(\frac{-1}{v}\right)[0, 1, 1, \cdots, 1].$$

(ii) From Lemma 3.2 (ii) and (3.8) it follows that

$$\mathbf{a}^{\vee}*\mathbf{a}^{\vee\sim}=\mathbf{a}^{\vee}*\mathbf{a}^{\sim\vee}=\left(\frac{-1}{v}\right)\mathbf{a}^{\vee}*\mathbf{a}^{\vee}=\left(\frac{-1}{v}\right)\sqrt{v}\,(\mathbf{a}\circ\mathbf{a})^{\vee}=\left(\frac{-1}{v}\right)\sqrt{v}\,[0,1,1,\cdots,1]^{\vee}$$

$$=\left(\frac{-1}{v}\right)[v-1,-1,-1,\cdots,-1].$$

This together with the proof of (i) shows that $\sqrt{(-1/v)}\,\mathbf{a}^{\vee}$ is a special binary coded pulser. The proof for $\sqrt{(-1/v)}\,\mathbf{a}^{\wedge}$ is similar and hence will be omitted. □

Remark. Hence, for a special binary coded pulser $\mathbf{a}$ of order p(= an odd prime), $P(\mathbf{a}, \omega^t)^2$ behaves like the square of the Gauss sum [15, pp. 207–208] (cf. [1, Thm. 7, pp. 349–353; Problems 13–16, p. 355] and [14, pp. 197–218]).

THEOREM 3.5. *If v = an odd prime p, then there is a unique normalized binary coded pulser of order p and its terms are given by the Legendre symbols* $(0/p)$, $(1/p)$, $(2/p), \cdots, (p-1/p)$.

Proof. Let $\zeta = e^{2\pi i/p}$.

(i) *Uniqueness.* Suppose $\mathbf{a}=[0, 1, a_2, \cdots, a_{p-1}]$ and $\mathbf{b}=[0, 1, b_2, \cdots, b_{p-1}]$ are normalized binary coded pulsers. Then Lemma 3.4 (i) implies that $P(\mathbf{a}, \zeta)=\varepsilon P(\mathbf{b}, \zeta)$, $\varepsilon=\pm 1$. However, the set $\{\zeta, \zeta^2, \cdots, \zeta^{p-1}\}$ is a basis for the field $\mathbb{Q}(\zeta)$ over the field of rational numbers $\mathbb{Q}$. So we have

$$1-\varepsilon=0, \quad a_2-\varepsilon b_2=0, \quad \cdots, \quad a_{p-1}-\varepsilon b_{p-1}=0.$$

Consequently $\varepsilon=1$ and $\mathbf{a}=\mathbf{b}$.

(ii) *Existence.* Let $\mathbf{c}=[c_0, c_1, c_2, \cdots, c_{p-1}]$ be given by the Legendre symbols; i.e., $c_j=(j/p)$ for all j. It is well known that $(P(\mathbf{c}, \zeta^t))^2$, the square of the Gauss sum, has the value 0 if $t\equiv 0 \pmod p$ and the value $(-1/p)p$ otherwise [15, pp. 207–208]. This result is equivalent to $\mathbf{c}^{\vee}\circ\mathbf{c}^{\vee}=(-1/p)\,[0, 1, 1, \cdots, 1]$. Then, since $\mathbf{c}^{\sim}=(-1/p)\mathbf{c}$ and $\mathbf{c}^{\sim\vee}=\mathbf{c}^{\wedge}$, it follows that $\mathbf{c}^{\wedge}\circ\mathbf{c}^{\vee}=[0, 1, 1, \cdots, 1]$. Applying $\vee$ to the last equation and using (3.8) and $\mathbf{c}^{\vee\vee}=\mathbf{c}^{\sim}$, we have $\mathbf{c}*\mathbf{c}^{\sim}=[p-1, -1, -1, \cdots, -1]$. □

For any odd prime p, by Theorem 3.5, there is a unique normalized binary coded pulser $\mathbf{c}$ of order p which is given by the Legendre sequence. We denote the corresponding polynomial $P(\mathbf{c}, X)$ by $F_p(X)$, so that $F_p(X)=\sum_{j=0}^{p-1}(j/p)X^j$. As a result of Lemma 3.4 and Theorem 3.5, for $\zeta=e^{2\pi i/p}$,

$$F_p(\zeta^t)=\left(\frac{t}{p}\right)F_p(\zeta) \quad \text{for } t=0, 1, 2, \cdots$$

and

$$F_p(\zeta)=\pm\sqrt{\left(\frac{-1}{p}\right)p}.$$

Now we define, for each odd integer v, a polynomial $F_v(X)$ with similar properties. Let $v=p_1^{\alpha_1}p_2^{\alpha_2}\cdots p_r^{\alpha_r}$ be the prime factorization of v and let $q_j=v/p_j$. The definition of $F_v(X)$ is

$$F_v(X)=\prod_{j=1}^{r}[F_{p_j}(X^{q_j})]^{\alpha_j}. \tag{3.10}$$

The standard properties of the Jacobi symbol imply that $F_v(X)$ has integral coefficients and

$$F_v(\omega^t)=\left(\frac{t}{v}\right)F_v(\omega) \quad \text{for } t=0, 1, 2, \cdots, \tag{3.11}$$

$$F_v(\omega) = \pm\sqrt{\left(\frac{-1}{v}\right)v}, \tag{3.12}$$

where $\omega = e^{2\pi i/v}$.

PROPOSITION 3.6. *If v is not a square and v has at least two distinct prime divisors, then there is no binary coded pulser of order v.*

Proof. Assume that there is a special binary coded pulser $\mathbf{a}$ of order v. By hypothesis we have $v = d \cdot q$, where $d > 1$, $(d, q) = 1$, and $q = p^{2m+1}$ for an odd prime p. We use the notation: $\omega = e^{2\pi i/v}$, $\zeta_q = e^{2\pi i/q}$, $\zeta_d = e^{2\pi i/d}$ and $\Phi_q(X) = \prod_{0<t<q,(t,q)=1}(X - \zeta_q^t)$ is the cyclotomic polynomial of order q (hence of degree $\varphi(q)$). The polynomials $F_d(X)$ and $F_q(X)$ are defined by (3.10), and so $F_d(\zeta_d) = \pm\sqrt{(-1/d)d}$ and $F_q(\zeta_q) = \pm\sqrt{(-1/q)q}$ by (3.12). The polynomial $P(\mathbf{a}, X)$ is $\sum_{j=0}^{v-1} a_j X^j$, and so, by Lemma 3.4 (i), $P(\mathbf{a}, \zeta_d) = P(\mathbf{a}, \omega^q) = \pm\sqrt{(-1/v)v}$ since $d > 1$. Both $P(\mathbf{a}, X)$ and $F_d(X)$ have integral coefficients, hence $P(\mathbf{a}, \zeta_d)$ and $F_d(\zeta_d)$ are in $\mathbb{Q}(\zeta_d)$, and hence their quotient $\sqrt{(-1/q)q} = F_q(\zeta_q)$ is in $\mathbb{Q}(\zeta_d)$.

Now we use q to define a new polynomial $f(X)$ by $f(X) = \prod_{0<t<q,(t/q)=1}(X - \zeta_q^t)$, so that degree $(f(X)) = \frac{1}{2}\varphi(q)$ because $(t/q) = (t/p)^{2m+1} = (t/p)$. Moreover,

$$f(X) = \text{g.c.d.}\,(\Phi_q(X), F_q(X) - F_q(\zeta_q)) \in \mathbb{Q}(\zeta_d)[X]$$

if we take (3.11) and the result that $F_q(\zeta_q) \in \mathbb{Q}(\zeta_d)$ into account.

Thus

$$\begin{aligned}[\mathbb{Q}(\omega): \mathbb{Q}(\zeta_d)] &= [\mathbb{Q}(\omega): \mathbb{Q}(\omega^q)] = [\mathbb{Q}(\omega^q)(\omega^d): \mathbb{Q}(\omega^q)] \\ &= [\mathbb{Q}(\zeta_d)(\zeta_q): \mathbb{Q}(\zeta_d)] \leqq \text{degree}\,(f(X)) \\ &= \tfrac{1}{2}\varphi(q),\end{aligned}$$

since $(d, q) = 1$. But then

$$\varphi(d)\varphi(q) = \varphi(v) = [\mathbb{Q}(\omega): \mathbb{Q}] = [\mathbb{Q}(\omega): \mathbb{Q}(\zeta_d)][\mathbb{Q}(\zeta_d): \mathbb{Q}] \leqq \tfrac{1}{2}\varphi(q)\varphi(d),$$

a contradiction. Thus there is no special binary coded pulser of order v. □

PROPOSITION 3.7. *If v is a square, then there is no binary coded pulser of order v.*

Proof. Assume that there is a special binary coded pulser $\mathbf{a} = [a_0, a_1, a_2, \cdots, a_{v-1}]$ of order v. By hypothesis we have $v = b^2$ for some odd integer b. We claim that $a_t = a_d$ whenever $(t, v) = d$. To see this, we first note that $(-1/v) = (-1/b)^2 = 1$, and hence, by Lemma 3.4: (i), $P(\mathbf{a}, \omega^t) = \pm b$ or 0; (ii) $\mathbf{a}^{\vee} = (1/b)[P(\mathbf{a}, 1), P(\mathbf{a}, \omega), P(\mathbf{a}, \omega^2), \cdots, P(\mathbf{a}, \omega^{v-1})]$ is also a special binary coded pulser. If, for each $d|v$, we define $g_d(X) = P(\mathbf{a}, X) - P(\mathbf{a}, \omega^d)$, then $g_d(X) \in \mathbb{Z}[X]$, and $\Phi_{v/d}(X)|g_d(X)$ as $g_d(\omega^d) = 0$ and $\Phi_{v/d}(X)$ is the minimal polynomial of ω^d over $\mathbb{Q}$. But $\Phi_{v/d}(X)|g_d(X)$ means $g_d(\omega^t) = 0$ whenever $(t, v) = d$; i.e., $P(\mathbf{a}, \omega^t) = P(\mathbf{a}, \omega^d)$ whenever $(t, v) = d$, whence

$$a_t = a_d \quad \text{whenever } (t, v) = d$$

as $\mathbf{a}^{\vee\vee\vee\vee} = \mathbf{a}$. This proves our claim.

Consequently,

$$\pm b = P(\mathbf{a}, \omega) = \sum_{j=1}^{v-1} a_j \omega^j = \sum_{\substack{d|v \\ d \neq v}} a_d \left(\sum_{\substack{0<t<v \\ (t,v)=d}} \omega^t \right).$$

Now, the *Ramanujan sum* $\sum_{0<t<v,(t,v)=d} \omega^t$ = sum of the primitive (v/d)th roots of unity is known to have the value $\mu(v/d)$ [10, Thm. 19, p. 99; Problem 18, p. 109]

where μ is the Möbius function. Therefore, $\pm b = \sum_{d|v, d\neq v} a_d \mu(v/d)$. Taking absolute values we have $b \leqq \sum_{d|v, d\neq 1} |\mu(d)|$. Now, if $b = p_1^{\beta_1} p_2^{\beta_2} \cdots p_r^{\beta_r}$ is the prime factorization of b, we have $b \geqq 3^r$ and $\sum_{d|v, d\neq 1} |\mu(d)| =$ number of nonempty subsets of $\{p_1, p_2, \cdots, p_r\} = 2^r - 1$. The inequality $3^r \leqq 2^r - 1$ is a contradiction. Thus there is no special binary coded pulser of order v. □

THEOREM 3.8. *If there is a binary coded pulser* **a** *of order v, then v is a prime.*

Proof. Without loss of generality, we may assume that **a** is special. Then by Proposition 3.6 and Proposition 3.7, v must be an odd power of a prime p, say $v = p^{2m+1}$. We shall show that $m = 0$.

Let $P(\mathbf{a}, X)$, $F_p(X)$ be defined as before and let $\Phi_{p^j}(X)$ be the cyclotomic polynomial of order p^j (and hence of degree $\varphi(p^j)$) so that $\Phi_{p^j}(X) = \Phi_p(X^{p^{j-1}})$ for $j \geqq 1$. We claim that for each $j = 1, 2, \cdots, 2m+1$, there exists $\varepsilon_j = \pm 1$ such that

$$P(\mathbf{a}, X) \equiv \varepsilon_j p^m F_p(X^{p^{j-1}}) \pmod{\Phi_{p^j}(X)}.$$

To see this, we use the facts that $P(\mathbf{a}, \rho)^2 = (-1/v)v = p^{2m}(-1/p)p$ for all ρ such that $\rho^v = 1$ and $\rho \neq 1$ (Lemma 3.4(i)), that $F_p(\theta)^2 = (-1/p)p$ for all θ such that $\theta^p = 1$ and $\theta \neq 1$ (square of Gauss sum), and thus any zero of $\Phi_{p^j}(X)$ is a zero of

$$P(\mathbf{a}, X)^2 - p^{2m} F_p(X^{p^{j-1}})^2 = (P(\mathbf{a}, X) - p^m F_p(X^{p^{j-1}}))(P(\mathbf{a}, X) + p^m F_p(X^{p^{j-1}})).$$

But each of these has integral coefficients, and so $\Phi_{p^j}(X)$, which is irreducible over $\mathbb{Q}$, must divide one or the other of the factors. This establishes our claim.

By the Chinese remainder theorem,

$$\frac{\mathbb{Q}[X]}{\langle X^{p^{2m+1}} - 1\rangle} \cong \prod_{j=0}^{2m+1} \frac{\mathbb{Q}[X]}{\langle \Phi_{p^j}(X)\rangle},$$

and hence there exists a unique $P(X) \in \mathbb{Q}[X]$ such that

1) degree $(P(X)) < p^{2m+1}$;
2) $P(X) \equiv \varepsilon_j F_p(X^{p^{j-1}}) p^m \pmod{\Phi_{p^j}(X)}$ for $j = 1, 2, \cdots, 2m+1$;
3) $P(X) \equiv 0 \pmod{\Phi_1(X)}$.

Therefore $P(X) = P(\mathbf{a}, X)$.

On the other hand, there is another explicit (global) expression for $P(X)$ available. In fact, the polynomial defined by

$$Q(X) = \sum_{j=1}^{2m} \varepsilon_j F_p(X^{p^{j-1}}) \left(\prod_{l=j+1}^{2m+1} \frac{1}{p} \Phi_p(X^{p^{l-1}}) \right) p^m + \varepsilon_{2m+1} F_p(X^{p^{2m}}) p^m$$

has rational coefficients and satisfies all three requirements 1), 2) and 3) as well. This can be seen by using

$$F_p(X^{p^{l-1}}) \equiv 0 \pmod{\Phi_{p^j}(X)} \quad \text{and} \quad \frac{1}{p}\Phi_{p^l}(X) \equiv 1 \pmod{\Phi_{p^j}(X)}$$

whenever $2m+1 \geqq l \geqq j+1 \geqq 2$.

As a result, $Q(X) = P(X) = P(\mathbf{a}, X)$. The only term in $Q(X)$ with X to the first power occurs when $F_p(X)$ is used, and its coefficient is $\varepsilon_1 (1/p)^{2m} p^m = \varepsilon_1 p^{-m}$. The only way that this coefficient can be equal to $a_1 = \pm 1$ is for $m = 0$. □

REFERENCES

[1] Z. I. BOREVICH AND I. R. SHAFAREVICH, *Number Theory*, Academic Press, New York, 1967.
[2] R. H. BRUCK, *Difference sets in a finite group*, Trans. Amer. Math. Soc., 78 (1955), pp. 464–481.

[3] F. F.-c. CHIANG, *The Theory and Design of A New Electromechanical Pulser*, Master's Thesis in electrical engineering, Texas Tech University, Lubbock, Texas, August, 1976.
[4] SHIN-HWA CHING, *A New Mathematical Design of Electromechanical Pulsers*, Master's Thesis in electrical engineering, Texas Tech University, Lubbock, Texas, May, 1975.
[5] J. W. COOLEY, P. A. W. LEWIS AND P. D. WELCH, *The finite Fourier transform*, IEEE Trans. Audio Electroacoust, AU-17 (1969), pp. 77–85.
[6] J. P. CRAIG, M. O. HAGLER AND K. I. SELIN, *Electromechanical pulser investigation*, PR A-3-1029 proj #4506 contract P-30602-73-C-0170, Rome Air Development Center Procurement Division (PMA) Griffiss AFB, NY 13441, April 1974; AD 787 674, National Technical Information Service, Springfield, Virginia 22151.
[7] J. P. CRAIG AND R. SAEKS, *An electromechanical pulser generator*, Proc. IEEE International Pulsed Power Conference, 76CH1147-8 Region 5, Institute of Electrical and Electronics Engineers, 1976, pp. IIB7–1 to IIB7–4.
[8] G. N. GLASOE AND J. V. LEBACQZ, *Pulse Generators*, MIT Radiation Laboratory Series #5, Boston Technical Publishers, Lexington, MA, 1964.
[9] S. W. GOLOMB, *Shift Register Sequences*, Holden-Day, San Francisco, 1967.
[10] E. GROSSWALD, *The Theory of Numbers*, Macmillan, New York, 1966.
[11] M. HALL, JR., *Cyclic projective planes*, Duke Math. J., 14 (1947), pp. 1079–1090.
[12] M. HALL, JR. AND H. J. RYSER, *Cyclic incidence matrices*, Canad. J. Math., 3 (1951), pp. 495–502.
[13] J. B. KELLY, *A characteristic property of quadratic residues*, Proc. Amer. Math. Soc., 5 (1954), pp. 38–46.
[14] E. LANDAU, *Elementary Number Theory*, Chelsea, New York, 1958.
[15] S. LANG, *Algebra*, Addison–Wesley, Reading, MA, 1970.
[16] M. NEWMAN, *Multipliers of difference sets*, Canad. J. Math., 15 (1963), pp. 121–124.
[17] O. PERRON, *Bemerkungen über die Verteilung der quadratischen Reste*, Math. Z., 56 (1952), pp. 122–130.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 3, September 1981

0196-5212/81/0203-0007 $01.00/0

EQUILIBRIUM AND STABILITY IN A PERIODIC MARKETING RING*

A. H. ZEMANIAN†

Abstract. A dynamic economic model for a periodic marketing network of the following sort is constructed herein. Traders operating out of urban centers pass among rural markets during the marketing week buying up a commodity, which they transport back to their respective urban centers and sell to wholesalers. The traders follow fixed and in general different, rings of markets, which they repeat from week to week. A set of nonlinear difference equations is devised that determines the time-dependent prices and commodity flows throughout the network. When the traders' rings are all the same, we prove that the system has a unique equilibrium state. This is established for two cases, one where the traders do not store goods and the second where they are allowed to do so. In the former case, we also establish conditions under which the equilibrium state is locally asymptotically stable. We finally show that, in the two-ring case, a peculiar price variation can occur, a result that may explain some of the reported "unpredictable" price swings in periodic markets.

Introduction. A common marketing system in the developing countries is the periodic marketing network [5]. It usually occurs in conjunction with other kinds of marketing and is just one part of the overall marketing system. Nonetheless, it is often the major part, especially in those third-world countries whose economies can be characterized as primarily rural, based on subsistence agriculture and having low effective demand for marketed goods.

Periodic markets can be described as follows. In a geographic region there appear many marketplaces, but only some of them are open as active markets on any single day. That is, most marketplaces become active markets on only one or perhaps several—but not on all—days of the week. Thus, on Monday a certain subset of the marketplaces open as markets, on Tuesday another subset of marketplaces open and so forth throughout the market week. This shifting pattern of active markets repeats itself from market week to market week.

There have been many anthropological, economic, geographic, historical and sociological studies of periodic markets [1], [2], but almost all of them have been nonmathematical, except perhaps for the statistical manipulation of field data. The comparatively few papers that do take a mathematical approach discuss either the graph-theoretic structure of periodic marketing networks including spatial and temporal patterns of trading, or aspects of central-place theory or static optimization ideas from economics [1], [2]. In a recent series of papers, we have undertaken the study of the dynamic economic behavior of periodic marketing systems [6], [7], [8], [9]. The present work is a continuation of that effort. Such a study is of interest because periodic marketing systems appear to react sluggishly to market disturbances and therefore are often in disequilibrium. Even when each market in the network achieves an equilibrium on each of its market days, it may well be that the overall marketing network remains in disequilibrium because it cannot react fast enough to the changing supply and demand conditions under which it operates.

A problem concerning the study of periodic markets is the fact that they appear in many different forms; that is, there are a variety of temporal and spatial patterns of market trading. Moreover, a particular marketing system may exhibit a combination of several such patterns. For the purposes of mathematical modeling, the only tractable approach appears to be the separate analysis of each form, uncombined with any other

* Received by the editors December 8, 1980. This work was supported by the National Science Foundation under grant MCS 78-01992.

† Department of Electrical Engineering, State University of New York, Stony Brook, New York 11794.

variety. Once each form is understood, it is hoped that a comprehension of combined systems may then be synthesized.

This work is an economic analysis of the following form of periodic marketing network, one that has not been so studied before. We assume that itinerant traders pass among many rural markets during the marketing week buying up a particular commodity. At the beginning of the next week, the traders converge upon a few urban centers where they either sell their accumulated goods to wholesalers or, if prices in the wholesale markets are unfavorable, store all or part of those goods. Each trader follows his own trading route, which we refer to as his "ring".

We analyze each trader as a firm that supplies the service of transferring ownership of the commodity over space and time. From this we obtain a characterization of each trader by means of an excess-supply function for the commodity (rather than the said service) in each market in which he trades. Upon aggregating these functions over all traders in every market and equating the result to the aggregate excess-demand function of all the other agents in the market, we obtain a set of simultaneous nonlinear difference equations as our model for the overall periodic marketing system. These equations determine the dynamic variations of all the prices and commodity flows. In particular, they explicate how price disturbances propagate from market to market in a step-by-step fashion, as was pointed out by W. O. Jones [3].

Once a dynamic economic model is at hand, a natural question to ask is whether it has an equilibrium state. We prove that a single-ring system whose traders do not store goods has a unique equilibrium state. Moreover, that state is asymptotically stable when certain conditions on the elasticities of the supply and demand functions are satisfied. For a perishable staple food, these elasticity conditions are not likely to be satisfied, which leads to the conclusion that periodic marketing networks that are adequately represented by our model tend toward instability. We also establish the existence of a unique equilibrium state for the more general case where traders may store goods; the asymptotic stability of that state is presently an open question. Finally, we examine a two-ring system and show, by example, that a sudden rise in demand in a wholesale market can lead to a subsequent fall in price in a rural market two days later. This may be one possible explanation of certain price variations observed in periodic markets that have been labeled as erratic and unpredictable [4, p. 22].

2. The behavior of a trader in a rural market. Throughout this work we shall use the same notation as that of our prior papers [6], [7], [8], [9]. Our first task is to assume a reasonable behavior for the traders when buying goods in a rural market. We shall assume that each trader is a profit maximizing firm that supplies the service of transferring ownership of a commodity over space and time. He does so by buying the commodity in a sequence of rural markets and selling it in an urban wholesale market. This cycle lasts for one market week, and the trader repeats his ring during each new market week. These ideas lead to an excess-supply function[1] for each trader in each rural market. Its derivation was presented in our prior works. To avoid repetition, we shall merely state the result here and refer the reader to those works for its derivation; see especially [8, § 2].

We assume that every market week has n days, not counting the rest days on which all markets are closed. The days of any week are indexed consecutively by $s = 1, 2, \cdots, n$. When $s = n$, we set $s+1 = 1$, and when $s = 1$ we set $s-1 = n$. The integer

[1] The reason we have chosen to represent the behavior of the trader in a rural market by an excess-supply function instead of an excess-demand function is that we wish to maintain notational conformity with our prior works on periodic markets.

t denotes the market day, and the integer ν is the index for the market week, also numbered consecutively. Thus, $t = \nu n + s$. Furthermore, we assume that every trader is in an urban center, on the first day of every week, where he sells or stores all of the goods he has accumulated during the preceding week. During the remaining $n - 1$ days of the week, he proceeds through a fixed route of rural markets, spending one day in each market.[2]

Let ϕ_{sj} represent the jth market that meets on the sth day of the marketing week. Thus ϕ_{1j} denotes an urban market, and ϕ_{sj}, where $2 \leqq s \leqq n$, denotes a rural market. Furthermore, we number all the traders in the entire marketing network by $i = 1, 2, \cdots, n$, and we attach i as a superscript to any symbol that pertains exclusively to the ith trader.

Figure 1 shows the excess-supply curve $S_s^i(p, t)$ that characterizes the ith trader, while he operates on day $t = \nu n + s$, in a rural market ϕ_{sj}, $2 \leqq s \leqq n$. As usual, p denotes the price of the commodity and q its quantity. $E_s^i(t)$ is the price the trader expects the

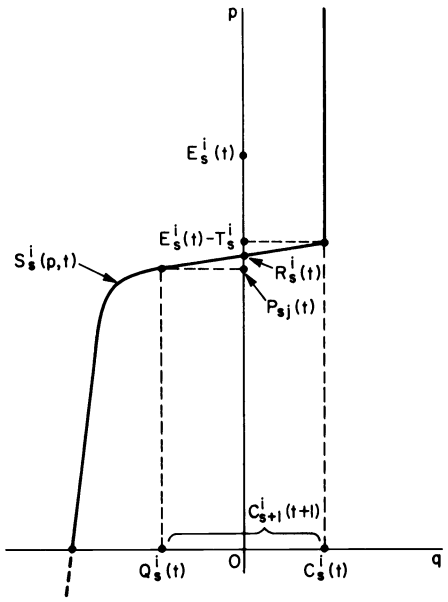


FIG. 1. *An excess-supply function for the* ith *trader in a rural market* ϕ_{sj}.

commodity will command when he returns to his urban market at the beginning of the next marketing week. It is determined by the trader's memory of past prices and his experience with price variations. Although we have been assuming in prior works that there is no market news, we could encompass a certain amount of news by specifying $E_s^i(t)$ as a function of prior prices, in not only the markets of his ring, but also in markets outside his ring. We do however maintain our assumption that no trader ever alters his ring.

$C_s^i(t)$ is the amount of goods the trader brings into ϕ_{sj} from the preceding market in his ring. Since he sells or stores all the goods he has on hand whenever he is in his urban market, we have $C_2^i(t) = 0$. The value T_s^i is a measure of the cost to the trader

[2] Actually, we allow a trader to visit the same marketplace on two or more days of a single week; we simply denote markets that meet in the same marketplace on different week days as being different markets.

in transferring one unit of the commodity from a seller in ϕ_{sj} to a buyer in his urban market at the beginning of the next market week. When the market price $P_{sj}(t)$ is larger than $E_s^i(t) - T_s^i$, the trader sells in ϕ_{sj} all of the amount $C_s^i(t)$, hence the vertical segment at the upper end of his excess-supply curve. When $P_{sj}(t)$ is considerably less than $E_s^i(t) - T_s^i$, the trader buys enough goods to fill or perhaps somewhat overfill his transporting capacity, hence the almost vertical portion of the lower end of his excess-supply curve. In between these two parts there is a transition region where the trader buys or sells some goods, but not enough to either fill or empty his transporting equipment. As was mentioned above, the reasons for this shape of a trader's excess-supply function are given in [8, § 2]. We merely add at this point that, by the conditions assumed later on, the traders of our model will normally buy—not sell—the commodity in any rural market. That is, the price $P_{sj}(t)$ will almost always remain below each trader's cross-over price $R_s^i(t)$. Moreover, except possibly for the rural markets that meet on the last day of the market week, $P_{sj}(t)$ will ordinarily lie only slightly below $R_s^i(t)$, so that the quantity $|Q_s^i(t)|$ bought by the trader adds to the amount he has already acquired but does not fill up his transporting capacity. This is the situation illustrated in Fig. 1.

3. The behavior of a trader in an urban market. We have yet to specify the economic behavior of the traders in the urban markets. We shall discuss two cases. In the first one, we assume that every trader sells all the goods $C_1^i(t)$ he brought into his urban market on the first day of each marketing week. Thus, on the axes of Fig. 1, his supply function is simply a vertical line that intersects the abscissa at $q = C_1^i(t) \geqq 0$. We shall refer to this case as the *no-storage model.*

The second case, which we refer to as the *storage model,* allows each trader to store goods in his urban center. The amount $A^i(t)$ he stores depends upon the current price $P_{1j}(t)$ in the urban market and the price $E_1^i(t)$ he expects therein one week hence. The trader's supply schedule $S_1^i(p, t)$ for this case was also derived in a prior work [7, § 4]. Again, we will not repeat those arguments but will merely describe their conclusion.

$S_1^i(p, t)$ is illustrated in Fig. 2. It lies in the first quadrant and terminates at the point where it meets the abscissa. When $P_{1j}(t) \geqq E_1^i(t)$, the trader sells all of the goods $G^i(t)$ he has on hand; $G^i(t)$ consists of the goods $C_1^i(t)$ he has just brought in from his last trip and the goods $A^i(t-n)$ in storage during the past week. This yields the upper vertical portion of $S_{1j}(p, t)$. As $P_{1j}(t)$ begins decreasing below $E_1^i(t)$, he begins storing goods for the coming week. Assume that the per-unit cost $Z^i(A^i(t))$ to the trader of storing goods increases continuously and strictly monotonically as the total amount $A^i(t)$ he stores increases up to his storage capacity B^i. As a consequence, he stores just that amount $A^i(t)$, where the per-unit cost of storing equals $E_1^i(t) - P_{1j}(t)$, so long as he does not exceed his storage capacity B^i. This gives the curved portion of $S_1^i(p, t)$. If he does fill B^i, he sells any goods on hand in excess of B^i for whatever price he can get. This accounts for the lower vertical portion of the supply curve of Fig. 2.

The illustration in Fig. 2 assumes that $E_1^i(t) - I^i$, where $I^i = Z^i(B^i)$ and $G^i(t) - B^i$ are both positive, so that the lower corner point lies within the first quadrant. There are two other possibilities. One is that $S_1^i(p, t)$ may meet the ordinate on its curved portion. At this point, $G^i(t) = A^i(t) \leqq B^i$, and the trader stores all his goods. For still lower prices, the trader continues to store all of $G^i(t)$, and $S_1^i(p, t)$ coincides with the ordinate down to the origin. As the second possibility, $S_1^i(p, t)$ may meet (and terminate at) the abscissa on its curved portion. At this point, $P_{1j}(t) = 0$, and the cost $Z^i(A^i(t))$ of storing goods equals the traders expected price $E_1^i(t)$; thus, the trader stores $A^i(t)$ and gives or throws away any additional goods he may have.

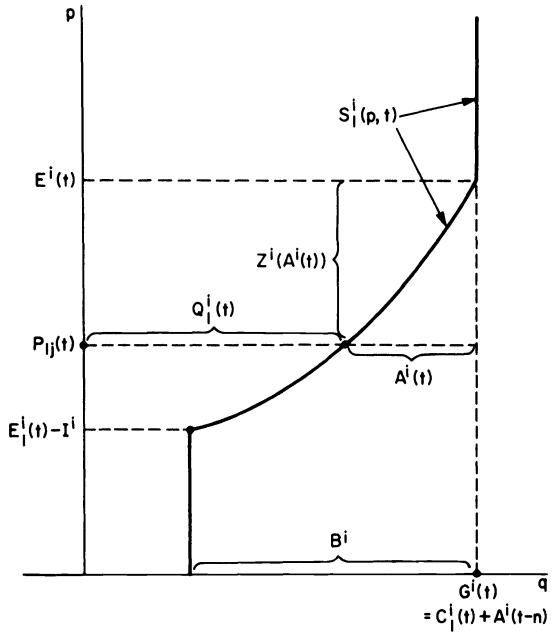


FIG. 2. *A supply function for the* ith *trader in an urban market* ϕ_{1j} *in the case of the storage model.*

We should point out that the storage behavior assumed herein is different from that of [7] in one respect. The price corresponding to the upper corner point in Fig. 2 is the trader's expected price $E_1^i(t)$, whereas in [7] it was his cost of buying and bringing one unit of the commodity into the urban market. Our present model allows storage for speculative purposes, whereas in the prior model the trader stored goods only to cut losses. It turns out that the present storage model has a simpler equilibrium state, for, as we shall see, no trader perpetually stores goods under equilibrium. In the prior model the latter could happen.

4. The dynamic models. We wish to formulate equations that determine all the dynamic price and commodity-flow variations throughout our periodic marketing network. To do so, we must specify the supply and demand schedules in functional form.

For $2 \leqq s \leqq n$, we formulate the ith trader's excess-supply function on day $t = s + \nu n$ by

$$S_s^i(p,t) = C_s^i(t) - V_s^i[E_s^i(t) - T_s^i - p], \tag{4.1}$$

where

$$C_s^i(t) = \begin{cases} -Q_2^i(t-s+2) - Q_3(t-s+3) - \cdots - Q_{s-1}^i(t-1) & \text{for } s = 3, 4, \cdots, n, \\ 0 & \text{for } s = 2. \end{cases} \tag{4.2}$$

As is indicated in Fig. 1, $-Q_s^i(t)$ denotes the amount of goods the ith trader buys in his rural market ϕ_{sj} on day t. Furthermore, the function V_s^i is assumed to satisfy the following conditions.

Conditions I. For each i and each s, V_s^i is a continuous nonnegative function on the real line such that $V_s^i(x) = 0$ for $x \leqq 0$, $V_s^i(x)$ is strictly increasing for $0 \leqq x < \infty$, and $V_s^i(x)$ tends to a finite limit $V_s^i(\infty)$ as $x \to \infty$.

For the purposes of this paper we need not specify anything more about V_s^i, even though the arguments of our prior works dictate a steplike shape for V_s^i. Also, (4.1) allows the shape of V_s^i to vary as s varies, but it seems likely that such shape changes will be minor. With respect to the axes of Fig. 1, the major variations will be horizontal shifts of S_s^i as $C_s^i(t)$ changes and vertical shifts of S_s^i as $E_s^i(t)$ and T_s^i change.

For $s=1$, every trader is in an urban market ϕ_{1j}. For the no-storage model, his supply function in ϕ_{1j} is perfectly inelastic

$$S_1^i(p,t)=C_1^i(t); \tag{4.3}$$

$C_1^i(t)$ is the amount of goods he has acquired on his last trip through the rural markets, namely

$$C_1^i(t)=-Q_2^i(t-n+1)-\cdots-Q_n^i(t-1). \tag{4.4}$$

For the storage model, however, a rather more complicated expression is needed to represent the three sections of $S_1^i(p,t)$ shown in Fig. 2.

$$S_1^i(p,t)=\begin{cases} G^i(t) & \text{for } p\geqq E_1^i(t), \\ \max\{0, G^i(t)-W^i[E_1^i(t)-p]\} & \text{for } E_1^i(t)-I^i\leqq p\leqq E_1^i(t), \\ \max\{0, G^i(t)-B^i\} & \text{for } p\leqq E_1^i(t)-I^i, \end{cases} \tag{4.5}$$

where $G^i(t)=C_1^i(t)+A^i(t-n)$. All the symbols herein have been defined in the preceding section except for W^i, which is the inverse mapping of the ith trader's storage-cost schedule Z^i. Z^i is assumed to satisfy the following conditions.

Conditions II. For each i, Z^i is a continuous, strictly increasing function on $0\leqq q\leqq B^i$. Moreover, $Z^i(0)=0$ and $Z^i(B^i)=I^i$.

The aggregate supply or excess-supply function of all the traders in a given market ϕ_{sj} is

$$S_{sj}(p,t)=\sum_{i}{}_{sj} S_s^i(p,t), \qquad 1\leqq s\leqq n, \tag{4.6}$$

where $\sum_{sj}$ denotes the sum over all indices i for those traders that operate in ϕ_{sj} on weekday s. The agents in ϕ_{sj} other than the traders are represented by an aggregate demand or excess-demand function $D_{sj}(p,t)$, which we take to be exogenously given. We also assume that in each urban market ϕ_{1j} those other agents only buy goods so that $D_{1j}(p,t)$ is positive for all $p>0$. However, in any rural market ϕ_{sj}, where $2\leqq s\leqq n$, we assume that there are both suppliers and local consumers, in addition to the traders. In view of this, we take $D_{sj}(p,t)$ to be negative for the usual values of p, but possibly positive for sufficiently small $p>0$. An aggregate demand curve for $s=1$ is shown in Fig. 3, and an aggregate excess-demand curve for $2\leqq s\leqq n$ is shown in Fig. 4. To be more precise, we impose the following conditions.

Conditions III. For every $t=\nu n+1$, where $\nu=0,1,2,\cdots$, and for every j in the index set of the urban markets ϕ_{1j}, $D_{1j}(p,t)$ is a positive, continuous, strictly decreasing function of p for $0<p<\infty$, such that $D_{1j}(p,t)\to\infty$ as $p\to 0+$ and $D_{1j}(p,t)\to 0+$ as $p\to\infty$. For every $t=\nu n+s$, where $\nu=0,1,2,\cdots$ and $s=2,\cdots,n$, and for every j in the index set of the rural markets ϕ_{sj}, $D_{sj}(p,t)$ is a continuous, strictly decreasing function of p for $0<p<\infty$, such that $D_{sj}(p,t)\to Q_{sj}^*\geqq 0$ as $p\to 0+$ and $D_{sj}(p,t)\to-\infty$ as $p\to\infty$. Also, for $p=0$, $D_{sj}(0,t)$ denotes the semiaxis $Q_{sj}^*\leqq q<\infty$.

The market-clearance conditions are obtained by equating supply to demand in each urban market and excess supply to excess demand in each rural market. That is,

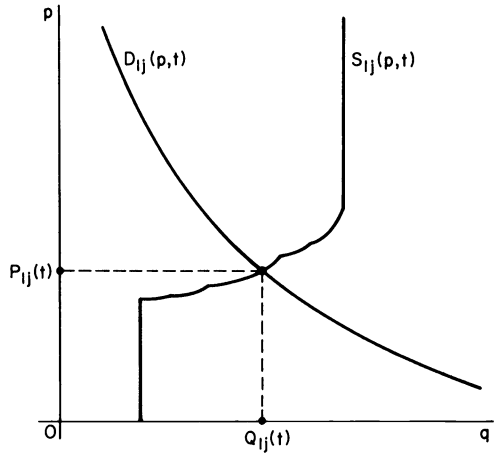


FIG. 3. *Aggregate supply $S_{1j}(p, t)$ and aggregate demand $D_{1j}(p, t)$ in an urban market ϕ_{1j}.*

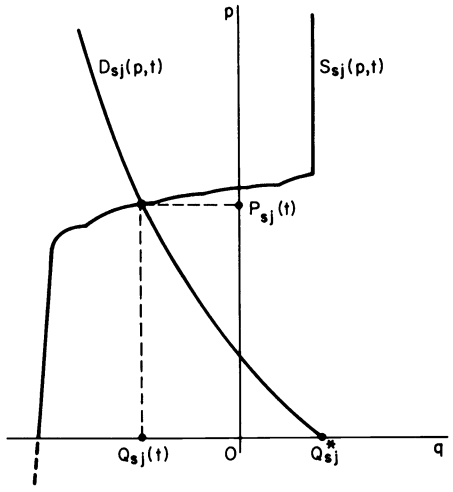


FIG. 4. *Aggregate excess supply $S_{sj}(p, t)$ and aggregate excess demand $D_{sj}(p, t)$ in a rural market ϕ_{sj}.*

for every j and t,

$$S_{1j}(p, t) = D_{1j}(p, t) \tag{4.7}$$

and, for every $s \geqq 2$, j and t,

$$S_{sj}(p, t) = D_{sj}(p, t), \qquad s = 2, \cdots, n. \tag{4.8}$$

To complete our models, we have to specify every trader's $E^i_s(t)$, the price he expects to find the next time he returns to his urban market—this expectation being held while he operates in his rural market ϕ_{sj} on day t. This can be done by specifying $E^i_s(t)$ through a memory function M^i_s of the prior prices in the marketing network. For the purposes of this paper, there is no need to be explicit at this point, but two reasonable conditions that we do impose on the M^i_s, are the following.

Conditions IV. The range value of each M^i_s increases or stays constant as any one or more of the arguments of M^i_s are increased, the remaining arguments being held

fixed. Also, if all prior prices in the trader's urban market are held constant at P_0, then M_s^i yields the same value P_0 once again for the next expected price.

We shall assume henceforth that Conditions I through IV are satisfied.

5. Recursive computations. With either the no-storage model or the storage model, we can recursively compute time series in every price and every commodity flow in the marketing system, once an appropriate set of initial conditions are assumed. For the storage model, assume that the amount stored $A^i(1)$ by the ith trader at $t=1$ for the forthcoming week is given for every i. Moreover, for both models, assume as given those prices prior to $t=1$ required in the arguments of all the memory functions M_s^i for the recursive computation of the $E_s^i(t)$ for $t=1, 2, 3, \cdots$. Finally, recall that for both models each trader's ring is taken to be known and fixed, that all the T^i are given, that all the aggregate demand and excess-demand functions $D_{sj}(p, t)$ are given exogenously for all $t \geqq 1$, and that $C_2^i(\nu n+2)=0$ for all i and ν.

We can now compute, at $t=2$, all the $E_2^i(2)$ from the memory functions M_2^i. Since $C_2^i(2)=0$, $S_2^i(p, t)$ is thereby determined with its upper vertical portion (see Fig. 1) lying on the ordinate. Upon aggregating over all the traders in a given market ϕ_{2j}, we obtain $S_{2j}(p, 2)$ for that market. The clearance equation (4.8) then yields the price $p=P_{2j}(2)$, in ϕ_{2j} at $t=2$, as indicated in Fig. 3. Referring back to Fig. 1, we obtain $Q_2^i(2)=S_2^i(P_{2j}(2), 2)$ and $C_3^i(3)=-Q_2^i(2)$.

Next, for $t=3$, we compute $E_3^i(3)$ from the memory functions M_3^i. Since $C_3^i(3)$ is known, we can repeat the computation of the preceding paragraph to get $P_{3j}(3)$, $Q_3^i(3)=S_3^i(P_{3j}(3), 3)$ and $C_4^i(4)=-Q_3^i(3)-Q_2^i(2)$. Continuing in this fashion along the days of the first marketing week (i.e., for $\nu=0$), we obtain all the prices for that week, as well as $C_1^i(n+1)$, from (4.4).

Now, for the no-storage model, we can aggregate over all the traders in ϕ_{1j} to obtain the perfectly inelastic supply function

$$S_{1j}(p, n+1)=\sum_{i\,1j} C_1^i(n+1).$$

By (4.7), we now get the price $p=P_{1j}(n+1)$. Of course, the amount $Q_1^i(n+1)$ sold by the ith trader in his ϕ_{1j} is equal to $C_1^i(n+1)$ in this case.

For the storage model, the memory functions serve again to give us $E_1^i(n+1)$. Also, the $A^i(1)$ are given as initial conditions, and this determines $G^i(n+1)=C_1^i(n+1)+A^i(1)$. Thus, we have fixed $S_1^i(p, n+1)$ for every i; see Fig. 2. Upon aggregating over the traders in any given urban market ϕ_{1j} to get $S_{1j}(p, n+1)$ and then using (4.7) as indicated in Fig. 3, we obtain the price $P_{1j}(n+1)$. Next, we refer again to Fig. 2 to obtain in addition $A^i(n+1)$ and $Q_1^i(n+1)$.

We can now repeat the computations of the second and third paragraphs of this section to get all the prices for the second week (i.e., for $\nu=1$), as well as $C_1^i(2n+1)$. This yields in turn $P_{1j}(2n+1)$ and $Q_1^i(2n+1)$ and, in the case of the storage model, $A^i(2n+1)$ as well.

This procedure can be continued to get as many prices or commodity flows as one may wish. We have in short constructed a no-storage model and also a storage model for our periodic marketing network, either of which completely determines a dynamic behavior.

6. The equilibrium state. Assume that all the exogenously given demand functions do not vary with t. An *equilibrium state* or an *equilibrium point* is a set of constant prices, one for each market, and also, in the case of the storage model, a set of constant stored amounts, one for each trader, such that the recursive analysis yields those prices

again and thereby constant time series in all prices and quantity flows. In other words, in an equilibrium state the dynamic response does not vary with time. (We shall show that the stored amounts are all zero in an equilibrium state of the storage model.)

Do our two dynamic models possess one or more equilibrium states? We have not been able to resolve this question for the general multi-ring models of the preceding section. However, when the marketing system consists of only one ring, the answer is "yes". More specifically, when there is one urban market and $n-1$ rural markets and when exactly one market opens on any given market day so that all traders follow the same ring, then each of our two models possesses a unique equilibrium state.

Since everything is constant with respect to t in an equilibrium state, we shall denote the time-invariant quantities by dropping their arguments in t.[3] Moreover, when there is only one ring, we drop the subscript j that—in the multi-ring case—numbers the markets open on a given market day. Thus, for example, for a single-ring system in an equilibrium state $P_{sj}(t)$ is replaced by P_s and $D_{sj}(p, t)$ is replaced by $D_s(p)$. Moreover, we let ϕ_s, where $s = 1, \cdots, n$, be the one and only market that opens on the sth weekday. Thus, ϕ_1 is the single urban market, and $\phi_2, \cdots, \phi_n$ are rural markets.

Our first theorem concerns the no-storage model.

THEOREM 1. *Assume that the no-storage model of a single-ring periodic marketing system satisfies the conditions of* § 4. *Assume furthermore that its demand functions do not vary with t. Then that model has a unique equilibrium state.*

Proof. We use a constructive proof. Our argument is illustrated in Fig. 5 for the case where $n = 4$. Since in an equilibrium state every price is constant with respect to

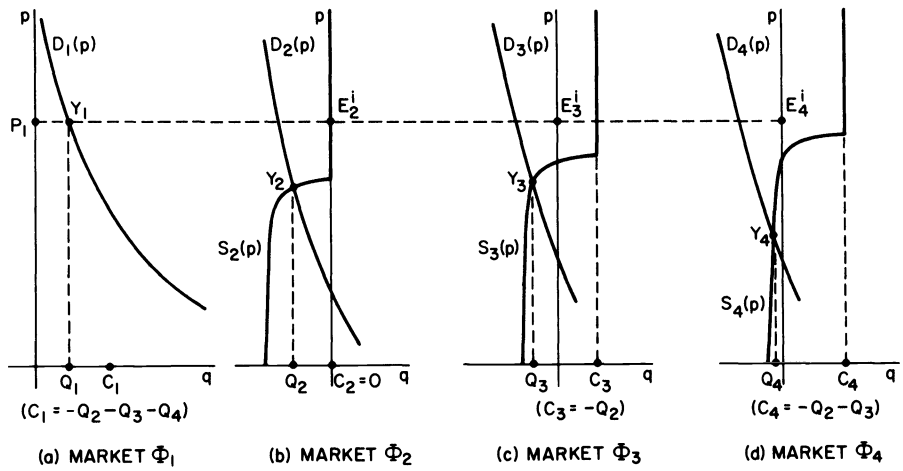


FIG. 5. *Illustration for the proof of Theorem* 1. S_s *and* D_s *are aggregate curves.*

t, we have $E_s^i = P_1$ for every $s = 2, \cdots, n$ and every i, according to Conditions IV. If we can find a value for $P_1 = E_s^i$ for which the amount Q_1 sold by the traders, as determined by clearance in ϕ_1, is equal to the amount C_1 brought into ϕ_1 by the traders, as determined by clearance in $\phi_2, \cdots, \phi_n$, then we will have found an equilibrium state. This is because the recursive computation of § 5 will then yield time-invariant prices

[3] To be sure, this is an abuse of notation, for in the time-varying case P_{sj} denotes a function that maps the integer t into a price whereas in the equilibrium case P_{sj} is a price. But, this notation is convenient and sufficiently clear.

and commodity flows everywhere. (The illustration in Fig. 5 shows a case where $Q_1 < C_1$. This of course is not a dynamic state; it is merely a mathematical construction.)

Recall that in ϕ_s, where $2 \leqq s \leqq n$, S_s shifts vertically downward as E_s^i is decreased, and S_s shifts horizontally to the left as C_s is decreased. In view of the conditions imposed on our supply and demand functions, we can assert the following: In ϕ_1, if P_1 is decreased continuously from ∞ to 0, then Q_1 increases continuously and monotonically from 0 to ∞.

On the other hand, in ϕ_2, by decreasing E_2^i from ∞ to 0, we increase Q_2 continuously and monotonically from $-\sum_i V_2^i(\infty)$ to 0. Furthermore, in ϕ_3, as we decrease E_3^i from ∞ to 0 and simultaneously decrease $C_3 = -Q_2$ from $\sum_i V_2^i(\infty)$ to 0, we move the intersection point Y_3 continuously along the S_3 curve toward its upper vertical portion. This means that $-Q_2 - Q_3$ decreases continuously and monotonically from $\sum V_3^i(\infty)$ to 0. The same variations hold in ϕ_4. We can therefore conclude that, as we decrease $P_1 = E_s^i$ from 0 to ∞, $C_1 = -Q_2 - Q_3 - Q_4$ decreases continuously and monotonically from a finite positive value to 0.

It now follows from the intermediate value theorem that there is a unique value of $P_1 > 0$ for which $Q_1 = C_1$. This ends the proof.

It is possible for a particular trader, say the ith trader, to be eliminated as an active agent under an equilibrium state. This occurs when the rural-market equilibrium prices P_s and his costs T_s^i are so high that

$$E_s^i - T_s^i - P_s = P_1 - T_s^i - P_s < 0$$

for every $s = 2, \cdots, n$. In this case, $C_1^i = 0$. As a result, the ith trader's supply function in ϕ_1 is simply the positive price axis. This trader neither acquires goods in the rural markets nor sells them in the urban market.

We turn now to the storage model. In this model, each trader has an expected price E_1^i in ϕ_1 as well as in the other markets. In an equilibrium state, because of the constancy of all prices and Conditions IV, we have that $P_1 = E_s^i$ for every $s = 1, \cdots, n$ and every i, as before. One consequence of this is that under equilibrium each trader operates in ϕ_1 at the upper corner point (E_1^i, G^i) of his supply function (see Fig. 2). This means that he never stores goods in an equilibrium state; he always sells whatever he has on hand in ϕ_1. Thus, $A^i = 0$ for all i. As in the no-storage model, if under equilibrium $C_1^i = 0$ for some i, then, since $A^i = 0$ too, the ith trader will be eliminated as an active agent.

Because every trader operates on (more precisely, just at the end of) the upper vertical portion of his supply curve when an equilibrium state exists, we can deduce the existence and uniqueness of such a state for the storage model directly from that of the no-storage model. Indeed, we can convert any storage model into a no-storage model simply by discarding every cost schedule Z^i and replacing the $S_1^i(p, t)$ of Fig. 2 by a vertical line that extends the upper vertical portion of $S_1^i(p, t)$ downward to the abscissa. It follows immediately that the storage model has an equilibrium state if and only if its associated no-storage model has one. Thus, we have the following corollary.

COROLLARY 1a. *The storage model of a single-ring periodic-marketing system that satisfies the hypothesis of Theorem* 1 *also has a unique equilibrium state.*

7. Asymptotic stability. As for the stability of the equilibrium state, we can show for the no-storage model, at least, that the equilibrium state is locally asymptotically stable, so long as a few more assumptions are made. This will be done by making a first-order Taylor expansion around the equilibrium state and then manipulating the resulting set of equations in differentials.

First of all, we maintain the assumption used in Theorem 1 that every demand or excess-demand function D_s does not depend upon t.

One assumption we add is that

$$E_s^i(t+s-1)=P_1(t), \qquad s=2,\cdots,n. \tag{7.1}$$

That is, every trader uses the last received price in ϕ_1 as his estimate of the next price in ϕ_1.

Another added assumption is that for all i and j,

$$T_s^i - T_s^j = T_\sigma^i - T_\sigma^j, \qquad s,\sigma=2,\cdots,n. \tag{7.2}$$

This means that the functions $p \mapsto V^i[P_1(t)-T_s^i-p]$ do not shift relative to each other as s varies. We may now set $T_s=\min_i T_s^i$ and

$$V[P_1(t)-T_s-p]=\sum_i V^i[P_1(t)-T_s^i-p]. \tag{7.3}$$

The function $p \mapsto V[P_1(t)-T_s-p]$ represents the aggregate excess-supply function of all the traders in ϕ_s. If T_s varies with s, the function shifts its position but does not change shape as s varies.

Finally, let $G_s=D_s^{-1}$ be the inverse function of D_s. Still, another assumption we now impose is that V and every G_s have continuous second derivatives. This means of course that the corner points of the V^i are assumed to be rounded off. Conditions I do not prohibit this.

We will need one more condition in order to establish asymptotic stability. It is rather complicated and is stated in Theorem 2 below.

To proceed, we can rewrite the clearance equations (4.7) and (4.8) as follows.

$$P_2(t+1)=G_2\{-V[P_1(t)-T_2-P_2(t+1)]\}, \tag{7.4}$$

$$\begin{aligned} P_s(t+s-1)=G_s\{&V[P_1(t)-T_{s-1}-P_{s-1}(t+s-2)] \\ &-V[P_1(t)-T_s-P_s(t+s-1)]\}, \qquad s=3,\cdots,n, \end{aligned} \tag{7.5}$$

$$P_1(t+n)=G_1\{V[P_1(t)-T_n-P_n(t+n-1)]\}. \tag{7.6}$$

These equations have the form $z=f(x, y, z)$, where x, y and z denote prices. (f does not depend on y in (7.4) and (7.6).) We denote equilibrium prices by x_0, y_0 and z_0. By the recursive analysis of § 5, each of (7.4)—(7.6) implicitly determines the price on the left-hand side in terms of prior prices. With z denoting that left-hand price, we may write $z=g(x, y)$ and $z_0=g(x_0, y_0)$. It follows from our differentiability assumptions that g is Lipschitz continuous so that

$$|z-z_0| \leqq K\|(x, y)-(x_0, y_0)\|, \tag{7.7}$$

where $\|\cdot\|$ denotes the Euclidean norm and K is a constant. By the twice differentiability of G_s and V, we have the following first-order Taylor expansion with remainder. Here, f_x^0, f_y^0 and f_z^0 denote equilibrium values of the first partial derivatives of f with respect to the indicated subscripts.

$$z=z_0+f_x^0(x-x_0)+f_y^0(y-y_0)+f_z^0(z-z_0)+r(x-x_0, y-y_0, z-z_0).$$

In view of (7.7) and the fact that $|r(\alpha, \beta, \gamma)|=o(\|(\alpha, \beta, \gamma)\|)$, this allows us to write the following equation in differentials around the equilibrium point, after we take $x \to x_0$ and $y \to y_0$.

$$dz=f_x^0\,dx+f_y^0\,dy+f_z^0\,dz. \tag{7.8}$$

We now apply (7.8) to (7.4), (7.5) and (7.6). Γ_s, where $s = 1, \cdots, n$, will denote the first derivative of G_s evaluated at the equilibrium value of its argument. Also, Ω_s, where $s = 2, \cdots, n$, will denote the derivative of V evaluated at the equilibrium value of $P_1(t) - T_s - P_s(t+s-1)$. By the chain rule for differentiation, we obtain

$$(1-\Gamma_2\Omega_2)\, dP_2(t+1) = -\Gamma_2\Omega_2\, dP_1(t), \tag{7.9}$$

$$\begin{aligned}\Gamma_s\Omega_{s-1}\, dP_{s-1}(t+s-2) + (1-\Gamma_s\Omega_s)\, dP_s(t+s-1) \\ = \Gamma_s(\Omega_{s-1}-\Omega_s)\, dP_1(t), \qquad s = 3, \cdots, n,\end{aligned} \tag{7.10}$$

$$dP_1(t+n) = \Gamma_1\Omega_n\, dP_1(t) - \Gamma_1\Omega_n\, dP_n(t+n-1). \tag{7.11}$$

Next, (7.9) is solved for $dP_2(t+1)$, which is then substituted into (7.10) for $s = 3$. The equation obtained is solved for $dP_3(t+2)$, which is in turn substituted into (7.10) for $s = 4$. Continuing this way, we get an expression for $dP_n(t+n-1)$ in terms of $dP_1(t)$. This is substituted into (7.11), which yields

$$dP_1(t+n) = H\, dP_1(t), \tag{7.12}$$

where

$$H = \Gamma_1\Omega_n - \frac{\Gamma_1\Gamma_n\Omega_n}{1-\Gamma_n\Omega_n}\left(\Omega_{n-1} - \Omega_n - \frac{\Gamma_{n-1}\Omega_{n-1}}{1-\Gamma_{n-1}\Omega_{n-1}}\right.$$
$$\left.\left(\Omega_{n-2} - \Omega_{n-1} - \frac{\Gamma_{n-2}\Omega_{n-2}}{1-\Gamma_{n-2}\Omega_{n-2}}\left(\cdots\left(\Omega_2 - \Omega_3 - \frac{\Gamma_2\Omega_2}{1-\Gamma_2\Omega_2}(-\Omega_2)\right)\right)\right)\right).$$

Consequently, we can state the following precise result, which holds under the assumptions imposed in § 4 and at the beginning of this section.

THEOREM 2. *The equilibrium state of the no-storage model of a single-ring periodic marketing system is locally asymptotically stable if* $|H| < 1$. *It is not locally asymptotically stable if* $|H| > 1$.

$|H|$ will clearly be less than 1 if the $|\Gamma_s|$ and $|\Omega_s|$ are small enough. We can therefore conclude with the following economic interpretations. The $|\Gamma_s|$ being small, means that the demand function D_1 of the wholesalers in the urban market and the excess-demand functions D_s of the suppliers and local consumers in the rural markets, are sufficiently elastic in the vicinity of the equilibrium prices. The $|\Omega_s|$ being small, means that the aggregate excess-supply functions of the traders in the rural markets are sufficiently inelastic in the vicinity of the equilibrium prices. Actually, if the good being traded is a perishable staple food, the condition on the $|\Gamma_s|$ is unlikely. D_1 and D_s are more commonly inelastic in this case. Moreover, the condition on the $|\Omega_s|$ also need not hold. As is indicated in Fig. 4, the intersection point is quite likely to be on the nearly horizontal part of the S_s curve. All this indicates that single-ring periodic marketing systems tend toward instability. This is another possible explanation of the observed, seemingly erratic, price behavior of periodic markets.

8. An "unexpected" price variation. We now show, by example, that, according to our model, a sudden rise in demand in an urban market can lead to a fall in price in a rural market two days later. This can be seen by examining a two-ring system with a three-day market week, shown in Fig. 6. ϕ_{11} and ϕ_{12} are urban markets, and ϕ_2 and ϕ_3 are rural markets. Each of the many traders follow the arrows around either the upper ring or the lower ring. For the sake of simplicity, we also assume that (7.1) and (7.2) hold, that the same number of traders operate out of ϕ_{11} as out of ϕ_{12}, that the V^i functions are all the same and that the traders do not store goods, even though none

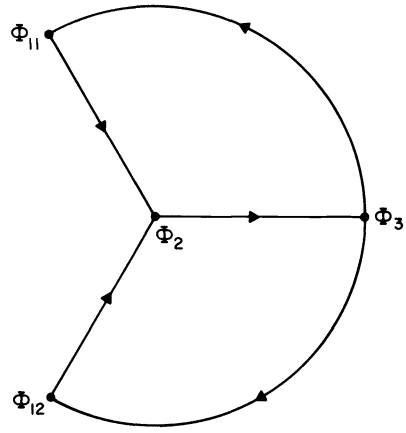


FIG. 6. *A two-ring periodic marketing network.*

of these assumptions are essential to our argument. It follows that the aggregate excess-supply curves for the two groups of traders have the same shape. It also follows that, if the demand functions D_{11} and D_{12} are identical, then the markets ϕ_{11} and ϕ_{12} are duplicates of each other—so far as supply and demand are concerned.

Let D_{12}, D_2 and D_3 be time-invariant. Consider first the case where $D_{11} = D_{12}$. We may combine ϕ_{11} and ϕ_{12} and conclude from Theorem 1 that our system has an equilibrium state. It is indicated by the solid-line curves of Fig. 7. Suppose that the system has operated in this equilibrium state for a while and then, on the urban-market day $t = 1$, the demand in ϕ_{11} jumps up to the curve D_{11}^*. By (7.1) the traders out of ϕ_{11}

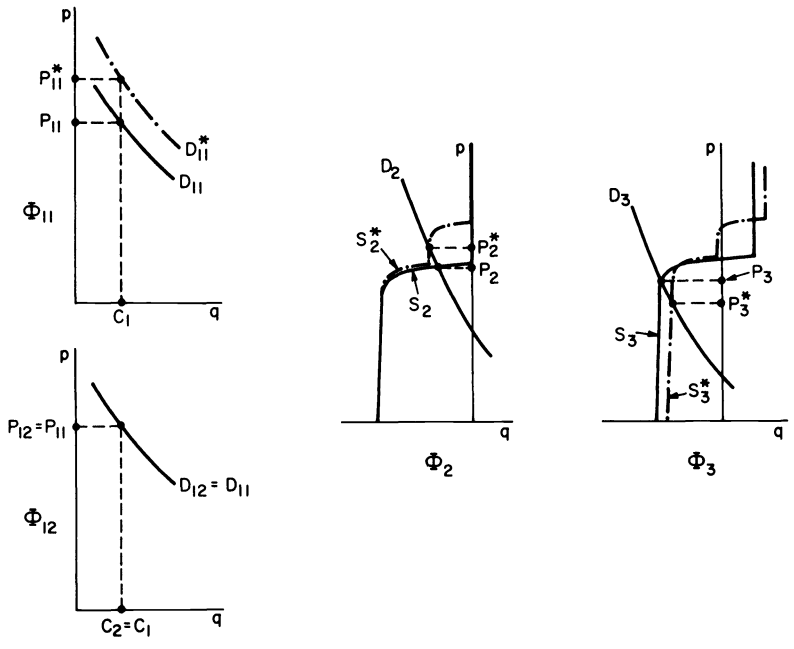


FIG. 7. *An example of an "unexpected" price variation.*

increase their expected price to P_{11}^* while the traders out of ϕ_{12} maintain their expected price at $P_{12} = P_{11}$. This raises the aggregate excess-supply curves for the traders out of ϕ_{11}, while the corresponding curves for the traders out of ϕ_{12} remain fixed. For ϕ_2 the sum of both curves is indicated by the dash–dot curve, and similarly for ϕ_3. If the D_2 and D_3 curves are positioned as indicated (i.e., relatively low supply in ϕ_2 and relatively high supply in ϕ_3), then the price in ϕ_2 rises to P_2^* on day $t = 2$, as expected, but the price in ϕ_3 falls to P_3^* on day $t = 3$.

This phenomenon can also be explained as follows. On day $t = 2$, the traders out of ϕ_{11} expect a high price P_{11}^* in ϕ_{11} at $t = 4$ and therefore bid the price up in ϕ_2. Moreover, they buy substantially more in ϕ_2 than they ordinarily did under the equilibrium state. On the other hand, the traders out of ϕ_{12} maintain their lower expected price P_{11} and, seeing an elevated price in ϕ_2, buy nothing in ϕ_2. Nevertheless, with D_2 positioned as indicated, the total amount of goods bought in ϕ_2 is significantly larger than it ordinarily was. This diminishes the total traders' demand in ϕ_3 for large supplies of the commodity. Since the supply curve $-D_3$ is large, the price in ϕ_3 drops to P_3^*.

REFERENCES

[1] R. J. BROMLEY, *Periodic Markets, Daily Markets and Fairs: A Bibliography*, Centre for Development Studies, University College of Swansea, Great Britain, 1974.

[2] ———, *Periodic Markets, Daily Markets and Fairs: A Bibliography Supplement to* 1979, Centre for Development Studies, University College of Swansea, Great Britain, 1979.

[3] WILLIAM O. JONES, *The structure of staple food marketing in Nigeria as revealed by price analysis*, Food Research Institute Studies, VIII (1968), pp. 95–123.

[4] ———, *Regional analysis and agricultural marketing research in tropical Africa: Concepts and experience*, Food Research Institute Studies, XIII (1974), pp. 3–28.

[5] ROBERT H. T. SMITH, *Periodic market-places and periodic marketing: Review and prospects—I and II*, Progress in Human Geography, 3 (1979), pp. 471–505; 4 (1980), pp. 1–31.

[6] A. H. ZEMANIAN, *Two-level periodic marketing networks without market news*, J. Math. Anal. Appl., 68 (1979), pp. 509–525.

[7] ———, *Two-level periodic marketing networks wherein traders store goods*, Geographical Analysis, 12 (1980), pp. 353–372.

[8] ———, *A dynamic economic model of periodic marketing rings*, Geographical Analysis, to appear.

[9] ———, *Economic models of periodic marketing systems*, *Proc.* 1980 IFIP *Conference on Global Modelling*, Dubrovnik, Yugoslavia, to appear.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 3, September 1981

0196-5212/81/0203-0008 $01.00/0

OPTIMAL FORMULAE OF THE CONDITIONAL MONTE CARLO*

B. L. GRANOVSKY†

Abstract. The conditional Monte Carlo method is designed for estimating a conditional expectation of a function, by sampling from an unconditional distribution. We give a description of the whole class of formulae of conditional Monte Carlo and on its basis derive an explicit expression for optimal (in the sense of dispersion) formulae. Such optimal formulae are constructed for three particular practical problems.

1. Introduction and summary. The conditional Monte Carlo method was discovered by Trotter and Tukey [1], [2]. The theory of the method was further developed in [3]–[5]. General discussion of the method and references can be found in [7]; particular mention should be made of the paper [6] devoted exclusively to the application of conditional Monte Carlo for solving the transport problem.

In § 2 of the present paper we give a description of the whole class of formulae of conditional Monte Carlo. From this description the formula proposed in the above papers appears as a particular case.

In § 3 we derive an explicit expression for optimal (in the sense of dispersion) formulae and in § 4 we construct such formulae for three particular problems considered in [1]–[6].

2. Formulae of conditional Monte Carlo. Conditional Monte Carlo is designed for estimating a conditional expectation of a function, by sampling from an unconditional distribution.

Let $\mathbf{z}$ be a random vector in Euclidean n-space R_n, having a probability density function (PDF) $h(\mathbf{z})$. Let $h(\mathbf{z}: \eta(\mathbf{z}) = \mathbf{y}_0)$ be a conditional PDF of $\mathbf{z}$ under the condition

$$\eta(\mathbf{z}) = \mathbf{y}_0$$

which is assumed to determine a surface S in R_n. Denote by F the set of functions φ having the conditional expectation

$$I(\mathbf{y}_0) = I(\varphi; \mathbf{y}_0) = E\{\varphi(\mathbf{z}); h(\mathbf{z}: \eta(\mathbf{z}) = \mathbf{y}_0)\}. \tag{1}$$

Conditional Monte Carlo consists of evaluating $I(\mathbf{y}_0)$ using the estimate $\hat{I}_N(\mathbf{y}_0)$ of the form

$$\hat{I}_N(\varphi; \mathbf{y}_0) = \hat{I}_N(\mathbf{y}_0) = N^{-1} \sum_{i=1}^{N} \varphi(T\mathbf{z}_i) W(\mathbf{z}_i). \tag{2}$$

Here $\mathbf{z}_i$, $i = 1, \cdots, n$ are independently sampled from the unconditional PDF $h(\mathbf{z})$, and the weight function W and the transformation T of R_n are chosen so as to provide the unbiasedness of $\hat{I}_N(\mathbf{y}_0)$, for all $\varphi \in F$:

$$E\hat{I}_N(\mathbf{y}_0) = I(\mathbf{y}_0), \qquad \varphi \in F. \tag{3}$$

As pointed out in the above-mentioned papers, the conditional Monte Carlo technique proves to be useful when the direct simulation of the conditional PDF would be a difficult process. In the given setting it is also worthwhile to note that the PDF $h(\mathbf{z})$ having the given conditional PDF $h(\mathbf{z}: \eta(\mathbf{z}) = \mathbf{y}_0)$ may not be prescribed, but chosen, for example, on the basis of convenience of simulation.

* Received by the editors May 12, 1980, and in final form December 26, 1980.
† Technion, Israel Institute of Technology, Haifa, Israel.

Throughout the paper any estimate (2) satisfying (3) will be called a formula of conditional Monte Carlo, and our first aim will be to give a description of all such formulae.

To proceed, observe first that from the above setting it follows immediately that condition (3) is equal to

$$E[\varphi(T\mathbf{z})W(\mathbf{z}); h(\mathbf{z})] = I(\varphi; \mathbf{y}_0), \qquad \varphi \in F. \tag{4}$$

Since $I(\varphi; \mathbf{y}_0)$ depends only on the values the function φ takes on S, the condition (4) implies that the transformation T should obey the condition

$$T\mathbf{z} \in S, \qquad \mathbf{z} \in R_n \qquad (\text{mod } h). \tag{5}$$

Choose further as in [3] a vector function $\mathbf{x} = \zeta(\mathbf{z})$ in such a way that the system

$$\mathbf{x} = \zeta(\mathbf{z}), \quad \mathbf{y} = \eta(\mathbf{z}), \quad \mathbf{z} \in R_n \tag{6}$$

establishes a 1 : 1 correspondence between the product space $X \times Y$ of pairs $(\mathbf{x}, \mathbf{y})$, and the space R_n. In other words $\mathbf{x}, \mathbf{y}$ provide a coordinate system in R_n, $\mathbf{x}$ being a system of coordinates in S. Denote by $\mathbf{z} = \nu(\mathbf{x}, \mathbf{y})$ the inverse transformation from $\mathbf{X} \times \mathbf{Y}$ to R_n, and suppose it has a Jacobian $J(\mathbf{x}, \mathbf{y}) = d\mathbf{z}/d\mathbf{x}\, d\mathbf{y}$. Then $f(\mathbf{x}, \mathbf{y}) = h(\nu(\mathbf{x}, \mathbf{y}))J(\mathbf{x}, \mathbf{y})$ is the PDF of $(\mathbf{x}, \mathbf{y})$, so that

$$h(\mathbf{z}\colon \eta(\mathbf{z}) = \mathbf{y}_0) = \frac{h(\nu(\mathbf{x}, \mathbf{y}_0))J(\mathbf{x}, \mathbf{y}_0)}{K_{\mathbf{y}}(\mathbf{y}_0)},$$

where $K_{\mathbf{y}}(\mathbf{y}) = \int_X f(\mathbf{x}, \mathbf{y})\, d\mathbf{x}$ is the PDF of $\mathbf{y}$.

Condition (5) on T can now be rewritten in the form $T\mathbf{z} = \nu(U\mathbf{x}, \mathbf{y}_0)$, where U is a transformation $X \to X$ and $\mathbf{x} = \zeta(\mathbf{z})$. So we obtain

$$I(\varphi; \mathbf{y}_0) = \int_X \varphi(\nu(\mathbf{x}, \mathbf{y}_0)) \frac{h(\nu(\mathbf{x}, \mathbf{y}_0))}{K_{\mathbf{y}}(\mathbf{y}_0)} J(\mathbf{x}, \mathbf{y}_0)\, d\mathbf{x} \tag{7}$$

and

$$\begin{aligned} E\hat{I}_N(\mathbf{y}_0) &= E[\varphi(T\mathbf{z})W(\mathbf{z}); h(\mathbf{z})] \\ &= \int_X \int_Y \varphi(\nu(U\mathbf{x}, \mathbf{y}_0))W(\nu(\mathbf{x}, \mathbf{y}))h(\nu(\mathbf{x}, \mathbf{y}))J(\mathbf{x}, \mathbf{y})\, d\mathbf{x}\, d\mathbf{y} \\ &= \int_X \varphi(\nu(U\mathbf{x}, \mathbf{y}_0)\, d\mathbf{x} \int_Y W(\nu(\mathbf{x}, \mathbf{y}))h(\nu(\mathbf{x}, \mathbf{y}))J(\mathbf{x}, \mathbf{y})\, d\mathbf{y}. \end{aligned} \tag{8}$$

In all the preceding papers U was chosen to be the identity. In what follows we will restrict ourselves to such a choice of U only for the sake of simplicity of exposition. An obvious modification of the resulting formula (9) below may easily be obtained by introducing the Jacobian of the transformation U.

Now from (7), (8) we derive that, under $U(\mathbf{x}) = \mathbf{x}$, the unbiasedness condition (4) implies the following necessary and sufficient condition on the weight function W:

$$\int_Y W(\nu(\mathbf{x}, \mathbf{y}))h(\nu(\mathbf{x}, \mathbf{y}))J(\mathbf{x}, \mathbf{y})\, d\mathbf{y} = \frac{h(\nu(\mathbf{x}, \mathbf{y}_0))}{K_{\mathbf{y}}(\mathbf{y}_0)} J(\mathbf{x}, \mathbf{y}_0), \qquad \mathbf{x} \in X. \tag{9}$$

The estimation procedure of $I(\mathbf{y}_0)$ can now be described in the following way. N values of $\mathbf{z}_i$, $i = 1, \cdots, N$ are sampled independently from $h(\mathbf{z})$, and the corresponding values of $\mathbf{x}_i, \mathbf{y}_i$, $i = 1, \cdots, N$ are calculated by formula (6). These values are used to

obtain the scores $\varphi(\nu(\mathbf{x}_i, \mathbf{y}_0))$ and the accompanying weights $W(\mathbf{z}_i) = W(\nu(\mathbf{x}_i, \mathbf{y}_i))$ satisfying (9).

It is easy to see that a partial solution of (9) is the following expression for W proposed in [1]–[6]:

$$W(\mathbf{z}) = W(\nu(\mathbf{x}, \mathbf{y})) = \frac{h(\nu(\mathbf{x}, \mathbf{y}_0))J(\mathbf{x}, \mathbf{y}_0)l(\mathbf{y})}{h(\nu(\mathbf{x}, \mathbf{y}))J(\mathbf{x}, \mathbf{y})K_{\mathbf{y}}(\mathbf{y}_0)}, \tag{10}$$

where $l(\mathbf{y})$ is an arbitrary PDF on Y.

3. Optimal formulae of conditional Monte Carlo. From (9) the problem naturally arises of determining the weight function $W = W^*$ minimizing the dispersion of a conditional Monte Carlo formula on the class of functions $\varphi \in F$.

We have

$$\operatorname{var} \hat{I}_N(\varphi; \mathbf{y}_0) = N^{-1}[E\{\varphi^2(T\mathbf{z})W^2(\mathbf{z})\} - I^2(\mathbf{y}_0)]$$

$$= N^{-1}\left[\int_X \varphi^2(\nu(\mathbf{x}, \mathbf{y}_0))\, d\mathbf{x} \int_Y W^2(\nu(\mathbf{x}, \mathbf{y}))h(\nu(\mathbf{x}, \mathbf{y}))J(\mathbf{x}, \mathbf{y})\, d\mathbf{y} - I^2(\mathbf{y}_0)\right].$$

Hence our problem is equivalent to the following one: Given h, ν find $W = W^*$ minimizing

$$R_W(\mathbf{x}) = \int_Y W^2(\nu(\mathbf{x}, \mathbf{y}))h(\nu(\mathbf{x}, \mathbf{y}))J(\mathbf{x}, \mathbf{y})\, d\mathbf{y}, \qquad \mathbf{x} \in X$$

under the condition (9). The solution is obtained immediately with the help of the Cauchy–Schwarz inequality

$$W^* = \frac{h(\nu(\mathbf{x}, \mathbf{y}_0))J(\mathbf{x}, \mathbf{y}_0)}{K_{\mathbf{y}}(\mathbf{y}_0)K_{\mathbf{x}}(\mathbf{x})}, \tag{11}$$

where $K_{\mathbf{x}}(\mathbf{x}) = \int_X h(\nu(\mathbf{x}, \mathbf{y}))J(\mathbf{x}, \mathbf{y})\, d\mathbf{y}$ is the PDF of $\mathbf{x}$ on X.

Formula (11) shows the optimal weight W^* and hence the optimal formula of conditional Monte Carlo should be independent of $\mathbf{y}$.

From (10) it also follows that if the transformation ζ is chosen so that the random variables $\mathbf{x} = \zeta(\mathbf{z})$ and $\mathbf{y} = \eta(\mathbf{z})$ are independent, then $W^* = 1$ and the optimal formula takes the form of the simplest Monte Carlo technique:

$$\hat{I}_N(\mathbf{y}_0) = N^{-1} \sum_{i=1}^{N} \varphi(\nu(\mathbf{x}_i, \mathbf{y}_0)).$$

Observe that in this case the optimal weight $W^* = 1$ can be obtained from (10) under $l(\mathbf{y}) = K_{\mathbf{y}}(\mathbf{y})$, but, in general, (10) does not reduce to (11) under any choice of $l(\mathbf{y})$.

4. Examples. In this section we obtain optimal formulae of conditional Monte Carlo for three particular practical problems considered in [1]–[6]. In each of these examples our formulae have less variance and are no more difficult to implement than the corresponding formulae in the above papers.

Example 1 ([4], [5]). Let $z \in (-\infty, +\infty)$ be a scalar random variable with a PDF $h(z)$. It is desired to estimate $E[\varphi(z)|z > 0]$. To present the problem in the form (1), put

$$y = \eta(z) = \begin{cases} 1, & z \geqq 0, \\ -1, & z < 0 \end{cases}$$

as in [4], [5] and choose $x = \zeta(z) = |z|$. Y-space in this case is a two element set $\{1; -1\}$, $y_0 = 1$, $X = (0, +\infty)$. The inverse transform is

$$z = \nu(x, y) = \begin{cases} x & \text{if } y = y_0 = 1, \\ -x & \text{if } y = -1, \end{cases}$$

and $J(x, y)$ is identically one.

Substituting these results in (10) we derive the optimal formula

$$\hat{I}_N(\varphi; y_0) = N^{-1} \sum_{i=1}^{N} \varphi(|z_i|) W^*(z_i),$$

where

$$W^*(z) = h(|z|)[p(h(-z) + h(z))]^{-1}$$

and

$$p = K_y(y_0) = \text{prob}\{z > 0\} = \int_0^{+\infty} h(x)\, dx.$$

The weight of the corresponding formula in [4], [5] is

$$W_\lambda(z) = \begin{cases} \lambda p^{-1}, & z > 0, \\ \dfrac{(1-\lambda)h(-z)}{ph(z)}, & z \leqq 0, \end{cases}$$

where λ is assumed to be a real number between 0 and 1. It is easy to check that

$$\text{var}[\varphi(|z|) W^*(z)] \leqq \text{var}[\varphi(|z|) W(z)], \qquad \varphi \in F,$$

with equality if and only if λ is not a constant but a function $\lambda^* = \lambda^*(z) = h(|z|)[h(z) + h(-z)]^{-1}$; in this case the two formulae are identical.

Example 2 ([1]–[3]). Let $\mathbf{z} = (z_1, \cdots, z_m)$ be a sample of m observations from the normal distribution $N(0, 1)$, $\eta(\mathbf{z}) = \max_i z_i - \min_i z_i$, the range of the sample and y_0 a prescribed positive number. Put $X = \{\mathbf{x} = (x_1, \cdots, x_m)\}$, $Y = \{y: y \geqq 0\}$ and consider the following transformation from X to $X \times Y$:

$$x_i = \frac{z_i}{\eta(\mathbf{z})}, \qquad i = 1, \cdots, m,$$

$$y = \eta(\mathbf{z}) = z_{(m)} - z_{(1)},$$

where $z_{(i)}$, $i = 1, \cdots, m$ are the order statistics of the sample. The transformation is the same as proposed in [1]–[3]. The inverse transform

$$\mathbf{z} = \nu(\mathbf{x}, y) = y\mathbf{x} \tag{12}$$

is uniquely determined in the field

$$y \geqq 0, \qquad \max_i x_i - \min_i x_i = 1.$$

The Jacobian of the transformation (12) is $J(\mathbf{x}, y) = y^{m-1}$.

Now, from (12) we have

$$f(\mathbf{x}, y) = (2\pi)^{-m/2} y^{m-1} \exp\left\{-\frac{y^2}{2} \sum_{i=1}^{m} x_i^2\right\}$$

and

$$K_{\mathbf{x}}(\mathbf{x}) = \int_0^{+\infty} f(\mathbf{x}, y)\, dy = 2^{-1}\pi^{-m/2}\Gamma\left(\frac{m}{2}\right)\left(\sum_{i=1}^{m} x_i^2\right)^{-m/2},$$

where $\Gamma(x)$ is the gamma-function.

With the help of (11) we come to the following expression for the optimal weight:

$$W^* = \frac{2\lambda^m \exp\left\{-\lambda^2 \sum_{i=1}^{m} z_i^2\right\}\left(\sum_{i=1}^{m} z_i^2\right)^{m/2}}{y_0\Gamma(m/2)K_y(y_0)},$$

where $K_y(y_0)$ is the value of the PDF of the range of the sample at the given point y_0, $\lambda = \lambda(y) = y_0/(y\sqrt{2})$. The value $K_y(y_0)$ can be determined from the known asymptotic expansion of the PDF $K_y(y)$ (see, e.g., [8]). The corresponding score is $\varphi((y_0/\eta(\mathbf{z}))\mathbf{z})$.

Example 3. One-dimensional transport problem ([5], [6]). Here we have to evaluate the integral of the form

$$I(z_0) = \int_{Z'} \varphi(\mathbf{z}', z_0)\, d\mathbf{z}',$$

where $\mathbf{z}' = (z_1, \cdots, z_n)$ and z_0 is a given real number. Denote $\mathbf{z} = (\mathbf{z}', z_{n+1})$ and let $h(\mathbf{z})$ be an arbitrary PDF on the $(n+1)$-dimensional space $Z = \{\mathbf{z}\}$. Now we can represent $I(z_0)$ in the form (1):

$$I(z_0) = \int_Z \tilde{\varphi}(\mathbf{z}) h(\mathbf{z}\colon \eta(\mathbf{z}) = y_0)\, d\mathbf{z}, \tag{13}$$

where

$$\tilde{\varphi}(\mathbf{z}) = \frac{\varphi(\mathbf{z})}{h(\mathbf{z}\colon \eta(\mathbf{z}) = y_0)}$$

and $\eta(\mathbf{z})$, y_0 are chosen so that the equation $\eta(\mathbf{z}) = y_0$ is equivalent to $z_{n+1} = z_0$. Following [5], [6] we define the transformation from Z to $X \times Y$ by

$$y = \eta(\mathbf{z}) = \frac{z_0}{z_{n+1}}, \qquad \mathbf{x} = \zeta(\mathbf{z}) = y\mathbf{z}' = y(z_1, \cdots, z_n) = \frac{z_0}{z_{n+1}}(z_1, \cdots, z_n),$$

and take $y_0 = 1$.

Hence the inverse transform is

$$\mathbf{z} = \nu(\mathbf{x}, y) = \left(\frac{x_1}{y}, \cdots, \frac{x_n}{y}, \frac{z_0}{y}\right),$$

with the Jacobian

$$J(\mathbf{x}, y) = \frac{|z_0|}{|y|^{n+2}}.$$

Now from (11), (13) the weight of the optimal formula may be readily found to be

$$W^* = \frac{1}{K_{\mathbf{x}}\left(\frac{z_0}{z_{n+1}}\mathbf{z}'\right)},$$

and the score is $\varphi((z_0/z_{n+1})\mathbf{z})$. From this it follows that the optimal formula in this case

has the form of an importance sampling technique, so that the PDF $h(\mathbf{z})$ can be adjusted for reduction of standard error.

Observe that in our notation the weight function W of the corresponding formula in [5], [6] is $W = 1/f(x/y)$ and the score is the same. So the two formulae coincide if the PDF $h(\mathbf{z})$ is chosen in such a way that $\mathbf{x}$ and y are statistically independent.

REFERENCES

[1] H. F. TROTTER AND J. W. TUKEY, *Conditional Monte Carlo for normal samples*, Symposium on Monte Carlo Methods, University of Florida, 1954, H. A. Meyer, ed., John Wiley, New York, 1956, pp. 64–79.

[2] H. J. ARNOLD, B. D. BUCHER, H. F. TROTTER AND J. W. TUKEY, *Monte Carlo techniques in a complex problem about normal samples*, ibid, pp. 80–88.

[3] J. M. HAMMERSLEY, *Conditional Monte Carlo*, J. Assoc. Comput. Mach., 3 (1956), pp. 73–76.

[4] J. G. WENDEL, *Groups and conditional Monte Carlo*, Ann. Math. Statist, 28 (1957), pp. 1048–1052.

[5] A. DUBI AND Y. S. HOROWITZ, *The interpretation of conditional Monte Carlo as a form of importance sampling*, SIAM J. Appl. Math., 36 (1979), pp. 115–122.

[6] D. W. DRAWBAUGH, *On the solution of transport problems by conditional Monte Carlo*, Nuclear Sci. Engng., 9 (1967), pp. 185–197.

[7] J. H. HALTON, *A retrospective and prospective survey of the Monte Carlo method*, SIAM Rev. 12 (1970), pp. 1–63.

[8] H. A. DAVID, *Order Statistics*, Wiley, New York, 1970.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 3, September 1981

0196-5212/81/0203-0009 $01.00/0

SEMIANTICHAINS AND UNICHAIN COVERINGS IN DIRECT PRODUCTS OF PARTIAL ORDERS*

DOUGLAS B. WEST† AND CRAIG A. TOVEY‡

Abstract. We conjecture a generalization of Dilworth's theorem to direct products of partial orders. In particular, we conjecture that the largest "semiantichain" and the smallest "unichain covering" have the same size. We consider a special class of semiantichains and unichain coverings and determine when equality holds for them. This conjecture implies the existence of k-saturated partitions. A stronger conjecture, for which we also prove a special case, implies the Greene–Kleitman result on simultaneous k- and $(k+1)$-saturated partitions.

1. Duality between semiantichains and unichain coverings. In this paper we study the relationship between semiantichains and unichain coverings in direct products of partial orders. Semiantichains are more general objects than antichains, and unichains are a restricted class of chains. The study of antichains (collections of pairwise unrelated elements) in partially ordered sets admits two approaches. The earlier arises from Sperner's theorem [32], which characterizes the maximum-sized antichains of a Boolean algebra. In general, Sperner theory obtains explicit values for the maximum size of antichains in partially ordered sets having special properties, and explicit descriptions of their composition. When the poset is ranked and the maximum-sized antichain consists of the rank with most elements, the poset has the *Sperner property*. Generalizations of Sperner's theorem have mostly consisted of showing that various posets have the Sperner property or stronger versions of the Sperner property. Greene and Kleitman [13] have given an excellent survey of results of this type.

Dilworth's theorem [4] bounds the size of the largest antichain by another invariant of the partial order. In particular, covering the partial order by chains is a "dual" minimization problem. No chain hits two elements of an antichain, so a covering always requires more items than any antichain has. Dilworth's theorem asserts that in fact the optimum sizes are always equal. The result does not give the extremal value or extremal collections, but it applies to all partially ordered sets. Generalizations of Dilworth's theorem have flowed less freely. A number of alternate proofs have been given, e.g. [3], [10], but the only broad extension we have is Greene and Kleitman's result [12] on k-families and k-saturated partitions.

The study of k-families began with Erdös. A k-family in a partially ordered set is a collection of elements which contains no chain of size $k+1$. An antichain is a 1-family. Erdös [6] generalized Sperner's theorem by showing that the largest k-family in a Boolean algebra consists (uniquely) of the k largest ranks. A (ranked) partial order satisfying this for all k is said to have the "strong Sperner property." Again, further Sperner-type results on k-families can be found in [13]. Clearly any chain contains at most k elements of a k-family, so any partition C of a partial order into chains $\{C_i\}$ gives an upper bound of $m_k(C) = \sum_i \min\{k, |C_i|\}$ on the size of the largest k-family. If the largest k-family has this size, the partition is called k-saturated. Greene and Kleitman proved there always exists a k-saturated partition, which for $k=1$ reduces to Dilworth's theorem. They showed further that for any k there exists a partition which

* Received by the editors September 12, 1980.

† Mathematics Department, Princeton University, Princeton, New Jersey 08544. The research of this author while at Stanford University was supported in part by the National Science Foundation under grant MCS-77-23738, and by the U.S. Office of Naval Research under contract N00014-76-C-0688. Reproduction in whole or in part is permitted for any purpose of the United States government.

‡ Operations Research Department, Stanford University, Stanford, California 94305.

is simultaneously k- and $(k+1)$-saturated. They applied lattice methods generalizing Dilworth's less well-known result [5] about the lattice behavior of antichains. Saks [30] gave a shorter proof of the existence of k-saturated partitions of P by examining the direct product of P with a k-element chain.

We consider a generalization of the Dilworth-type idea of saturated partitions to the direct product of any two partial orders. Sperner theory has also discussed direct products. A *semiantichain* in a direct product is a collection of elements no two of which are related if they are identical in either component. The class of semiantichains includes the class of antichains. If the largest semiantichain still consists of a single rank, then the direct product has the *two-part Sperner property*. Results of this nature have been proved by Katona [21], [23], Kleitman [24] and Griggs [15], [17], with extensions to k-families by Katona [22], Schonheim [31] and recently by Proctor, Saks, and Sturtevant [27]. Examples where maximum-sized semiantichains are not antichains were examined by West and Kleitman [33] and G. W. Peck [26].

To generalize Dilworth's theorem to semiantichains we need a dual covering problem. Semiantichains are more general objects than antichains, so we need more restricted objects than chains. We define a *unichain* (one-dimensional chain) in a direct product to be a chain in which one component remains fixed. Alternatively it is the product of an element from one order with a chain from the other. Two elements on a unichain are called *unicomparable*. Clearly no semiantichain can contain two elements of a unichain, so the largest semiantichain is bounded by the smallest covering by unichains. After [33], West and Saks conjectured that equality always holds. We have not proved equality for general direct products, but we prove a special case here. Also, we make a stronger conjecture analogous to Greene and Kleitman's simultaneous k- and $(k+1)$-saturation. If one of the partial orders is a chain of $k+1$ elements, the conjecture reduces to their result.

Note that maximizing semiantichains and minimizing unichain coverings are dual integer programs. One such formulation has as constraint matrix the incidence matrix between elements and unichains. Showing that the underlying linear program has an integral optimal solution would prove the conjecture, by guaranteeing that the integer program has no "duality gap."

These dual programs form an example of the frequent duality between "packing" problems and "covering" problems (see [1], [2], [7], [8], [11], [19], [25], [29]). Dilworth's theorem is another example; Dantzig and Hoffman [3] deduced it from duality principles. Hoffman and Schwartz [20] also used integer programming ideas to prove a slight generalization of Greene and Kleitman's k-saturation result by transforming the problem into a transportation problem. These methods work partly because any subset of a partial order is still a partial order. However, a subset of a direct product need not be a direct product. Indeed, subsets of direct product orders frequently have duality gaps between their largest semiantichains and smallest unichain coverings. (The smallest example is a particular 7-element subset of the product of a 2-element chain with a 3-element chain.)

Dilworth's theorem can also be proved by transforming it to a bipartite matching problem or a network flow problem (see [9], [10]). The difficulty in applying these latter methods to direct products is that unicomparability, unlike comparability, is not transitive. Much is known about the integrality of optima when the constraint matrix is totally unimodular, balanced, etc., as summarized by Hoffman [18]. Unfortunately, none of the several integer programming formulations we know of for this direct product problem have any of those properties. Finally, Greene and Kleitman use lattice theoretic methods because the set of k-families and maximum k-families form well-

behaved lattices. We have found no reasonable partial order on semiantichains or maximum semiantichains.

In the case where the largest semiantichain is also an antichain, network flow methods can be used to prove the conjecture. This result will appear in a subsequent paper. In § 2 we find necessary and sufficient conditions for equality to hold when semiantichains and unichain coverings are required to have a particularly nice property called "decomposability." When this happens, the size of the optimum is determined by the sizes of the largest k-families in the two components. In § 3 we develop the stronger form of the conjecture and show it holds in this case. We note with boundless ambition that if the first conjecture is true we can begin to ask about the existence of "k-saturated partitions" of direct products into unichains, analogous to k-saturated partitions of posets.

Before embarking on the subject of decomposability, we note that this duality question can be phrased as a problem in graph theory. The "comparability graph" of a partially ordered set is formed by letting (x, y) be an edge in $G(P)$ if x is related to y in P. An antichain becomes an independent set of vertices; a chain becomes a complete subgraph. Dilworth's theorem states that the independence number $\alpha(G)$ equals the clique covering number $\theta(G)$. When we take direct products, the "unicomparability graph" is just the product graph $G(P)\times G(Q)$.[1] Now independent sets are semiantichains and cliques are unichains, and again we want to show $\alpha = \theta$. Comparability graphs are perfect graphs, but it is not true in general for products of perfect graphs that $\alpha = \theta$. (Example:  , where the left factor is perfect, but not a comparability graph.) We can ask for what subclasses of perfect graphs does $\alpha(G\times H) = \theta(G\times H)$?

2. Decomposability. We consider semiantichains and unichain coverings which arise from partitions of the component orders. We will use $d(P, Q)$ to denote the size of the largest semiantichain in $P\times Q$.

Partition P and Q into collections of antichains $\mathscr{A}$ and $\mathscr{B}$. Any matching of antichains in $\mathscr{A}$ with antichains in $\mathscr{B}$ induces a semiantichain when the complete direct product of each matched pair is included. An antichain which can be formed in this way is called *decomposable.*

Given partitions of P and Q into antichains, it is a simple algebraic consequence that the largest decomposable semiantichain we can form from them is obtained by matching the largest from each, then the next largest, and so on. We call this the "greedy product" of two partitions, and its size is

$$g(\mathscr{A}, \mathscr{B}) = \sum_i |A_i||B_i|, \quad \text{where } A_i \geqq A_{i+1} \quad \text{and} \quad B_i \geqq B_{i+1}.$$

Now partition P and Q into collections of chains $\mathscr{C}$ and $\mathscr{D}$. This induces a unichain covering of $P\times Q$. For each pair (C_i, D_j), we cover the sub-product $C_i\times D_j$. It is easy to see we do this with fewest unichains if we take $\min\{|C_i|, |D_j|\}$ copies of the longer chain. Again, a unichain covering so formed is called a *decomposable* covering. Its size, a "pairwise minimum" function generalizing m_k, is

$$m(\mathscr{C}, \mathscr{D}) = \sum_{i,j} \min\{|C_i|, |D_j|\}.$$

[1] Independence number = size of largest set of mutually nonadjacent vertices. Clique covering number = size of smallest collection of complete subgraphs which together touch all vertices. Product graph $G\times H$ has as vertices the Cartesian product of the vertex sets of G and H. (u, v) and (u', v') are joined by an edge if $u = u'$ and (v, v') is an edge of H or $v = v'$ and (u, u') is an edge in G.

Using Greene and Kleitman's terminology, we let $d_k(P)$ denote the size of the largest k-family in P and put $\Delta_k(P) = d_k(P) - d_{k-1}(P)$. Let $\Delta^P \cdot \Delta^Q = \sum_k \Delta_k(P)\Delta_k(Q)$. (We note this is a quantity which appears independently in [29], where Saks proved $d_1(P \times Q) \leqq \Delta^P \cdot \Delta^Q$.)

To further simplify notation, let a_i and b_i be the size of the ith largest antichains in $\mathscr{A}$ and $\mathscr{B}$. Since $g(\mathscr{A}, \mathscr{B})$ depends only on the sizes in the partition, we will speak interchangeably of $g(\mathscr{A}, \mathscr{B})$ and $g(\langle a_i \rangle, \langle b_i \rangle)$ even if there is no decomposition corresponding to those numbers.

THEOREM 1. *For any antichain partitions $\mathscr{A}$ and $\mathscr{B}$ and chain partitions $\mathscr{C}$ and $\mathscr{D}$ of partial order P and Q,*

$$g(\mathscr{A}, \mathscr{B}) \leqq \Delta^P \cdot \Delta^Q \leqq m(\mathscr{C}, \mathscr{D}). \tag{0}$$

Furthermore, equality holds on the left if and only if

$$b_k > b_{k+1} \Rightarrow \sum_{i \leqq k} a_i = d_k(P), \tag{1}$$

$$a_k > a_{k+1} \Rightarrow \sum_{i \leqq k} b_i = d_k(Q), \tag{2}$$

$$b_k = b_{k+1} \quad \text{and} \quad a_k = a_{k+1} \Rightarrow \sum_{i \leqq k} a_i = d_k(P) \text{ or } \sum_{i \leqq k} b_i = d_k(Q). \tag{3}$$

Also, equality holds on the right if and only if

$$\Delta_k(P) > \Delta_{k+1}(P) \Rightarrow \mathscr{D} \text{ is } k\text{-saturated, and } \mathscr{C} \text{ is } l\text{-saturated whenever } \mathscr{D} \text{ has a chain of size } l. \tag{4}$$

Equality on the right is also equivalent to the statement obtained by exchanging $\mathscr{D}$ with $\mathscr{C}$ and Q for P in (4).

Proof. The first inequality holds by the same argument that made the greedy product the best way to match up antichains. Increasing a_k (beginning with $k = 1$, then 2, etc.) by shifting units from smaller a_i can only increase g, since those units will be paired with larger b_i than before. We must find an upper bound on this process.

The union of k antichains forms a k-family, so $\langle a_i \rangle$ is a nonincreasing sequence with $\sum_{i \leqq k} a_i \leqq d_k(P) = \sum_{1 \leqq i \leqq k} \Delta_k$ and similarly for b_i. So, we increase a_1 to $\Delta_1(P)$ and b_1 to $\Delta_1(P)$, then increase a_2 and b_2, etc., until $a_k = \Delta_k(P)$ and $b_k = \Delta_k(Q)$. It is important to note that $\Delta_k \geqq \Delta_{k+1}$, a nontrivial result proved in [12]. This guarantees that the nonincreasing character of the sequences will be preserved by the process. If we begin with an actual partition $(\mathscr{A}, \mathscr{B})$, we end with $\Delta(P) \cdot \Delta(Q)$ without decreasing the value of g.

When will equality hold? If $\langle a_i \rangle$, $\langle b_i \rangle$ are the sequences for $\mathscr{A}$ and $\mathscr{B}$ and $\sum_{i \leqq k} a_i$ is less than $d_k(P)$ for some k with $b_k > b_{k+1}$, we can increase g by increasing a_k at the expense of the smallest a_i. (Technically, we increase d_j for the smallest j such that $a_j = a_k$.) If $a_k = a_{k+1}$ and $b_k = b_{k+1}$, but both initial segments sum to less than the respective d_k, there will be room to gain by making such a change in both sequences simultaneously. On the other hand, if (1)–(3) are never violated, all the (legal) switches made to reach $\Delta^P \cdot \Delta^Q$ will leave them satisfied and produce no gain, so equality holds.

The second inequality is more subtle. We need more notation. Let $\alpha_k(\mathscr{C})$ be the number of chains in partition $\mathscr{C}$ which have at least k elements. If a partition of P is simultaneously $(k-1)$- and k-saturated, by definition $m_{k-1}(\mathscr{C}) = d_{k-1}(P)$ and $m_k(\mathscr{C}) = d_k(P)$. Subtracting the first from the second yields $\alpha_k(\mathscr{C}) = \Delta_k(P)$. So, if a completely saturated partition exists, the number of chains with exactly k elements will always be $\Delta_k(P) - \Delta_{k+1}(P)$. Let $\beta_k(P) = \Delta_k(P) - \Delta_{k+1}(P)$.

Next we cite the discrete analogue of integration by parts. Assuming the boundary terms vanish,

$$\sum_k u_k(v_k - v_{k+1}) = \sum_k (u_k - u_{k-1})v_k.$$

For u_k plug in d_k of one partial order, and for v_k use Δ_k of the other. Since $\beta_k = \Delta_k - \Delta_{k+1}$, we have

$$\sum d_k(P)\beta_k(Q) = \sum \Delta_k(P)\Delta_k(Q) = \sum \beta_k(P)d_k(Q). \tag{5}$$

By grouping pairs of chains appropriately, it is easy to see

$$\sum_i m_{|C_i|}(\mathscr{D}) = m(\mathscr{C}, \mathscr{D}) = \sum_j m_{|D_j|}(\mathscr{C}). \tag{6}$$

Now let $\mathscr{C}^*$ be a collection of chains with $\beta_k(P)$ of size k and $\mathscr{D}^*$ a collection with $\beta_k(Q)$ chains of size k. $\mathscr{C}^*$ and $\mathscr{D}^*$ may not exist as chain decompositions of P and Q, but as we did with antichains we can still apply the function m to those collections of chain sizes. In particular, $\mathscr{C}^*$ and $\mathscr{D}^*$ behave like completely saturated partitions, with $m_k(\mathscr{C}^*) = d_k(P)$ and $m_k(\mathscr{D}^*) = d_k(Q)$. Applying this to (6), we get

$$\sum d_k(P)\beta_k(Q) = m(\mathscr{C}^*, \mathscr{D}^*) = \sum \beta_k(P)d_k(Q). \tag{7}$$

When we use $\mathscr{D}$ rather than $\mathscr{D}^*$, the first half of (6) gives

$$m(\mathscr{C}^*, \mathscr{D}) = \sum \beta_k(P)m_k(\mathscr{D}) \geqq \sum \beta_k(P)d_k(Q), \tag{8}$$

with equality if and only if $\mathscr{D}$ is k-saturated whenever $\beta_k(P) > 0$, i.e., when $\Delta_k(P) > \Delta_{k+1}(P)$. Similarly, $m(\mathscr{C}, \mathscr{D}^*) \geqq m(\mathscr{C}^*, \mathscr{D}^*)$.

Now, if $\gamma_k(\mathscr{D})$ is the number of chains in $\mathscr{D}$ of size k, the other half of (6) gives

$$m(\mathscr{C}^*, \mathscr{D}) = \sum m_k(\mathscr{C}^*)\gamma_k(\mathscr{D}) = \sum d_k(P)\gamma_k(\mathscr{D}). \tag{9}$$

Replacing $\mathscr{C}^*$ by an actual partition $\mathscr{C}$ gives

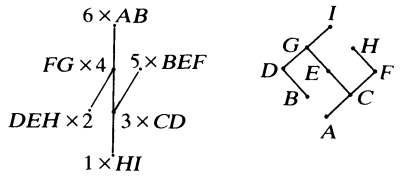
$$m(\mathscr{C}, \mathscr{D}) = \sum m_k(\mathscr{C})\gamma_k(\mathscr{D}) \geqq \sum d_k(P)\gamma_k(\mathscr{D}). \tag{10}$$

(5)–(10) combine to give

$$m(\mathscr{C}, \mathscr{D}) \geqq m(\mathscr{C}^*, \mathscr{D}) \geqq m(\mathscr{C}^*, \mathscr{D}^*) = \sum \Delta_k(P) \cdot \Delta_k(Q). \tag{11}$$

For equality to hold every step of the way, the conditions are as stated in the theorem, i.e., saturation requirements when β_k and γ_k are nonzero. Note that passing through $m(\mathscr{C}, \mathscr{D}^*)$ gives us the other set of conditions. The two are equivalent. □

Of course, if equality holds on both sides of (0) the desired duality holds. It has not been shown that the conditions for equality hold when the extremal semiantichain and unichain covering are both decomposable. Even if they do, the extremal packing and covering are not always decomposable, although there always exists a maximal decomposable semiantichain (i.e., no larger semiantichain contains it). Furthermore, the size of the optimal semiantichain and unichain covering may be strictly greater or strictly less than $\Delta^P \cdot \Delta^Q$. The first example of a direct product with no decomposable maximum-sized semiantichain was found by Saks [28]. Pictured in Fig. 1, it has $\Delta^P \cdot \Delta^Q = 13$, but the largest semiantichain has 14 elements, as indicated. The smallest example we know of is the product in Fig. 2a. The largest decomposable semiantichains have 9 elements, but it is not hard to find one of size 10, namely {1*a*, 1*b*, 1*c*, 2*d*, 2*e*, 2*f*, 3*d*, 3*e*, 3*b*, 3*c*}, indicated by large dots. Meanwhile, $m(21, 2211) = 10$. The unichain covering is indicated by heavy lines. However, when a slight change is made to reach



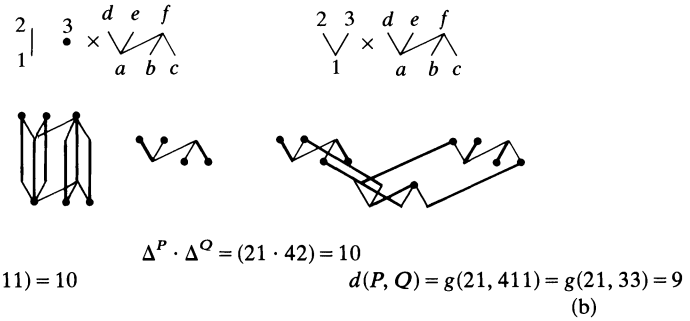
$g(2211 \times 32211) = 13 = \Delta(P) \cdot \Delta(Q)$,
$14 = m(33, 522) = m(411, 522) = d(P, Q)$

FIG. 1. "*little H*" × "*big H*".

Fig. 2b (adding the relation $3 > 1$), the semiantichain of size 10 disappears. Now the largest semiantichain is decomposable ($g(21, 411) = 9$), but the smallest unichain covering is not.

The usefulness of decomposable objects is that the extremal value among such objects can be computed quickly. For unichain coverings we can consider the broader class of quasi-decomposable coverings. These fix a partition of only one of the partial orders, then match each chain in that partition with some partition of the other order. We do best by providing a k-saturated partition for each k-chain. Then, if Q had the fixed partition, the size of the induced covering is $\sum d_k(P)\gamma_k(\mathscr{D})$. In the proof above, this is $m(\mathscr{C}^*, \mathscr{D})$, so such coverings are also bounded by $\Delta^P \cdot \Delta^Q$.

In this broader class less is required for equality. In particular, if one of the orders has a completely saturated partition, $\mathscr{D}$ becomes $\mathscr{D}^*$ and there is a quasi-decomposable unichain covering of size $\Delta^P \cdot \Delta^Q$. Unfortunately Fig. 2b shows that even when both P



$\Delta^P \cdot \Delta^Q = (21 \cdot 42) = 10$

$d(P, Q) = m(21, 2211) = 10$ (a)

$d(P, Q) = g(21, 411) = g(21, 33) = 9$ (b)

FIG. 2. *Nondecomposability.*

and Q have completely saturated partitions, there need not be a semiantichain of this size. Here duality still holds, though, because the minimum covering is not even quasi-decomposable, but is smaller yet. As with decomposable coverings, the optimum quasi-decomposable covering is easily computed. Not all chain partitions $\mathscr{D}$ of Q need be considered; chain partitions whose sequences are refinements of others are always dominated by the latter. In general, any covering by disjoint unichains can be expressed by partitioning the direct product into suitable subproducts such that the covering is the union of decomposable coverings of the subproducts. However, this formulation is unwieldy. Quasi-decomposable coverings give a quick near-optimal value which can help reduce the search for the optimal.

As for the usefulness of decomposability, we see that products of posets with completely saturated partitions will have unichain coverings of size $\Delta^P \cdot \Delta^Q$. Note also

that when the partial orders can be decomposed into antichains of sizes Δ_k, there will be a decomposable semiantichain of size $\Delta^P \cdot \Delta^Q$. This condition says the largest k-families are obtained by uniting the first k of some sequence of antichains. (This is not always true; in the poset of Fig. 3 no largest 2-family contains a largest 1-family.)

In particular, strongly Sperner posets satisfy the the latter condition. Sufficient to imply the strong Sperner property is the LYM property. (For a survey of results on LYM orders, see [13] once again.) The question of whether LYM orders always have completely saturated partitions remains open (see [14], [16]). If so, then products of LYM orders would have this "two-part Dilworth property." In any case equality certainly holds for products of symmetric or skew chain orders, etc., which are strongly Sperner and have completely saturated partitions.

3. Magic triples. We now discuss the analogue of a "simultaneously k- and $(k+1)$-saturated partition" for direct products.

We define a *magic triple*[2] $(\mathcal{S}, \mathcal{U}, x)$ in a direct product $P \times Q$ to be a maximum-sized semiantichain $\mathcal{S}$, a minimum-sized unichain covering $\mathcal{U}$, and an element x in P or Q satisfying the following properties.

1) x is the fixed element of unichains in $\mathcal{U}$ the same number of times it is a component of elements in $\mathcal{S}$.
2) When x is deleted, the restrictions of $\mathcal{S}$ and $\mathcal{U}$ to the smaller direct product are still extremal.

CONJECTURE. *A magic triple exists for every $P \times Q$, and hence the duality conjecture follows by induction on $|P|+|Q|$.*

Of course, if the duality conjecture is true in general, then property (2) above will hold whenever property (1) holds. Showing that implication holds without assuming the duality conjecture would make it easier to show magic triples always exist.

The magic triple conjecture is particularly satisfying because, although inductive, it is symmetric in P and Q. In their proof Greene and Kleitman had to consider two cases, corresponding to whether the element x belongs to P or to Q. The conjecture also explains the peculiarity in their result of guaranteeing simultaneous k- and $(k+1)$-saturation but being unable to guarantee more at one time. (The usual example that more cannot be guaranteed simultaneously is "little H" in Fig. 1.)

If Q is a $(k+1)$-chain, then any semiantichain in $P \times Q$ "projects down" to a $(k+1)$-family in P of the same size, since it uses $k+1$ disjoint antichains of P in the $k+1$ "copies" of P. Conversely, any $(k+1)$-family in P gives rise to (several) semiantichains of that size, so $d_{k+1}(P) = d(P, Q)$. A unichain covering of $P \times Q$ collapses to a partition $\mathcal{C}$ of P by collapsing the unichains that vary in Q to their fixed elements in P. Since Q can be covered by a single chain, such an element of P need not appear in any other unichain. If the unichain covering is minimal, the same chain decomposition of the remaining elements of P will be used in each of the $k+1$ copies of P in $P \times Q$, and all the P-unichains used will have at least $k+1$ elements. So, the bound $m_{k+1}(\mathcal{C})$ given by the corresponding partition $\mathcal{C}$ of P has the same size as the unichain covering.

Suppose magic triples exist, and hence duality holds. By the discussion above, the collapsed partition $\mathcal{C}$ is $(k+1)$-saturated. If the magic triple for $P \times Q$ has its "element" x in Q, then $\mathcal{C}$ is also k-saturated. If x is in P, we use induction on $|P|$. Obtaining a k and $(k+1)$-saturated partition and corresponding k and $(k+1)$-families for $P-x$, we add x to the families and as a single element chain to the partition. The properties

[2] Such a triple was originally called a "Catholic cucumber" due to late-night slurring of "the element is Q-crossed as many times as it is Q-covered."

of a magic triple guarantee the resulting partition of P is k and $(k+1)$-saturated, and the resulting collections are largest k and $(k+1)$-families.

Note that we have required a triple. It may be that for any extremal semiantichain or unichain covering there exists an example of the other that with it will form a triple. However, it is not true that any pair $(\mathcal{S}, \mathcal{U})$ will extend to a triple. For example, when the partial order of Fig. 3 is crossed with itself, there are (among others) two largest semiantichains and two largest unichain coverings which extend to triples when paired correctly but not when paired the other way.

FIG. 3. *M over W.*

If a pair $(\mathcal{S}, \mathcal{U})$ admits a sequence of elements such that successive restrictions of this pair form magic triples until the partial orders are exhausted, we call them completely mutually saturated. Theorem 2 is a sufficient condition for complete mutual saturation which applies to products of partial orders satisfying the conditions for equality in Theorem 1. It would be nice to strengthen this theorem by removing the words "of the same size", i.e., to show that if a maximum-sized semiantichain and minimum-sized unichain covering are both decomposable, then they have the same size.

Also, we note that the converse of the theorem is false, as shown by the examples in Fig. 2. The $(\mathcal{S}, \mathcal{U})$ pairs shown are not decomposable, but they are completely mutually saturated. The correct sequence of elements to be eliminated starts with $\{3\}$ in Fig. 2a and with $\{1\}$ in Fig. 2b. Then the reduced pair $(\mathcal{S}', \mathcal{U}')$ (see proof below) are decomposable, and the theorem can be applied to complete the sequence.

THEOREM 2. *If a direct product order has a largest semiantichain and a smallest unichain covering of the same size which are both decomposable, then they are completely mutually saturated.*

Proof. The element chosen to complete the magic triple can be any element on the chain which is shortest of both partitions. Let $\mathcal{S}$ be the semiantichain (induced by $\mathcal{A}$ and $\mathcal{B}$), $\mathcal{U}$ the unichain covering (induced by $\mathcal{C}$ and $\mathcal{D}$), and assume $\mathcal{C}$ has the shortest chain so $x \in P$. Then we claim $\mathcal{D}$ must be 1-saturated, and x appears in some antichain paired with the maximum-sized antichain of Q in $\mathcal{S}$. We show this will make it a magic triple. When x is removed, what remains of $\mathcal{S}$ and $\mathcal{U}$ will be extremal and decomposable for $(P-x)\times Q$, so we can repeat this until the orders are exhausted.

Let $\mathcal{A}'$ and $\mathcal{C}'$ be the reduced antichain and chain partitions of $P-x$, and let $d(P, Q)$ denote the size of the largest semiantichain in $P\times Q$. $d(P-x, Q)$ is bounded from above by the reduced decomposable covering, which gives the first inequality below. The middle equality follows since x lies on the shortest chain. That is, when $\mathcal{C}$ and $\mathcal{D}$ induce a unichain covering, the elements on the shortest chain always appear as fixed elements crossed with a longer chain in the other order. Removing such an element removes from the count the number of chains in $\mathcal{D}$. So we have

$$d(P-x, Q) \leqq m(\mathcal{C}', \mathcal{D}) = m(\mathcal{C}, \mathcal{D}) - |\mathcal{D}| = |\mathcal{U}| - |\mathcal{D}|. \tag{12}$$

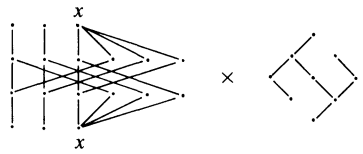
On the other hand, $d(P-x, Q)$ is bounded from below by the restriction of $\mathcal{S}$, giving the first inequality below. The second follows because in $\mathcal{A}\times\mathcal{B} = \mathcal{S}$, x must be paired with some antichain in $\mathcal{B}$, which has at most $d_1(Q)$ elements. Finally, since $\mathcal{D}$

is a partition it has at least $d_1(Q)$ chains, by Dilworth's theorem. This gives us

$$d(P-x, Q) \geqq g(\mathscr{A}', \mathscr{B}) \geqq g(\mathscr{A}, \mathscr{B}) - d_1(Q) \geqq |\mathscr{S}| - |\mathscr{D}|. \tag{13}$$

Since $|\mathscr{S}| = |\mathscr{U}|$, all the inequalities in (12) and (13) become equalities. In particular, $|\mathscr{D}| = d_1(Q)$, and $d_1(Q)$ is the size of the antichain matched with x's antichain. Also, $m(\mathscr{C}', \mathscr{D}) = g(\mathscr{A}', \mathscr{B})$, so $(\mathscr{S}, \mathscr{U}, x)$ is a magic triple, $\mathscr{S}'$ and $\mathscr{U}'$ are decomposable and equal and the argument can be applied to $(P-x)\times Q$ to complete the desired sequence of elements. □

We close with the only example we have yet found where neither the maximum semiantichain nor the minimum unichain covering is decomposable. One factor is the order "big H" devised by Saks and mentioned previously. The other is an example devised by Griggs [16] to show lack of implication among various poset properties. After much worry, we found the semiantichain and unichain coverings both of size 40 pictured in Fig. 4. Again the elements of the semiantichain appear as heavy dots, one



Fish × *Big H*

$d(P, Q) = 40$,
$\Delta \cdot \Delta = g(6633, 32211) = 39$,
$m(4442211, 522) = 42$,

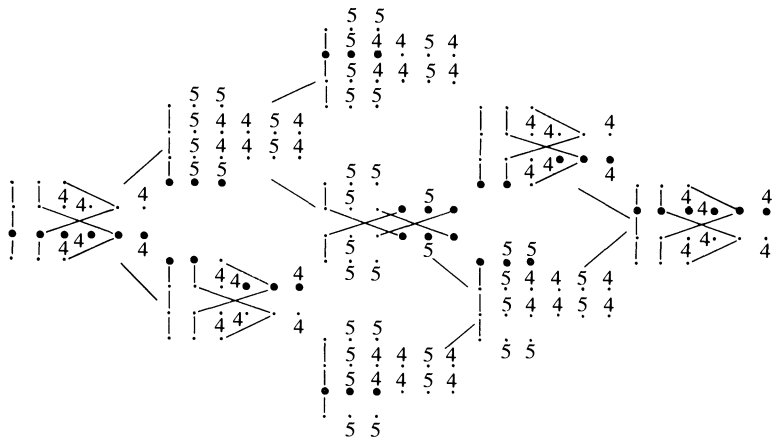


FIG. 4.

on each unichain. Elements labeled 4 or 5 in the direct product are covered by unichains which are copies of 4 or 5 element maximal chains in "big H". Although not decomposable, this pair still extends to a magic triple by selecting either of the two points of highest degree (marked x) in the "fish". After they are removed, the reduced semiantichain and unichain covering are still extremal but no longer extend to a magic triple. We are left with four disjoint products, including two copies of Saks' example and two selections of a 2-family from "big H". By choosing different extremal pairs, we can continue finding magic triples until the orders are exhausted.

REFERENCES

[1] V. CHVÁTAL, *On certain polytopes associated with graphs*, Centre de Recherches Mathématiques, 238, Université de Montréal, October, 1972.

[2] V. CHVÁTAL, *Edmonds polytopes and a hierarchy of combinatorial problems*, Discrete Math., 5 (1973), pp. 305–337.

[3] G. B. DANTZIG AND A. J. HOFFMAN, *Dilworth's theorem on partially ordered sets*, in Linear Inequalities and Related Systems, Annals of Mathematics Studies 38, H. W. Kuhn and A. W. Tucker, eds., Princeton University Press, Princeton, NJ, 1956, pp. 207–214.

[4] R. P. DILWORTH, *A decomposition theorem for partially ordered sets*, Ann. Math., 51 (1950), pp. 161–166.

[5] ———, *Some combinatorial problems on partially ordered sets*, in Combined Analysis, R. Bellman and M. Hall, eds., Proc. Symposium on Applied Mathematics, American Mathematical Society, Providence, RI, 1960, pp. 85–90.

[6] P. ERDÖS, *On a lemma of Littlewood and Offord*, Bull. Amer. Math. Soc., 51 (1945), pp. 898–902.

[7] J. EDMONDS, *Covers and packings in a family of sets*, Bull. Amer. Math. Soc., 68 (1962), pp. 494–499.

[8] ———, *Submodular functions, matroids, and certain polyhedra*, in Combinatorial Structures and Their Applications, Proc. Calgary International Conference, 1969, Gordon and Breach, New York, 1970, p. 69.

[9] L. R. FORD, JR., AND D. R. FULKERSON, *Flows in Networks*, Princeton University Press, 1962.

[10] D. R. FULKERSON, *Note on Dilworth's decomposition theorem for particlly ordered sets*, Proc. Amer. Math. Soc., 7 (1956), pp. 701–702.

[11] ———, *Anti-blocking polyhedra*, J. Combin. Theory, 12 (1972), pp. 50–71.

[12] C. GREENE AND D. J. KLEITMAN, *The structure of Sperner k-families*, J. Combin. Theory, 20 (1976) pp. 41–68.

[13] ———, *Proof techniques in the theory of finite sets*, in Studies in Combinatorics, G.-C. Rota, ed., Studies in Mathematics 17, Mathematical Association of America, 1978, pp. 22–79.

[14] J. R. GRIGGS, *Sufficient conditions for a symmetric chain order*, SIAM J. Applied Math., 32 (1977), pp. 807–809.

[15] ———, *Another three-part Sperner theorem*, Stud. Appl. Math., 58 (1977), pp. 181–184.

[16] ———, *On chains and Sperner k-families in ranked posets*, preprint.

[17] J. R. GRIGGS AND D. J. KLEITMAN, *A three-part Sperner theorem*, Discrete Math., 17 (1977), pp. 281–289.

[18] A. J. HOFFMAN, *The role of unimodularity in applying linear inequalities to combinatorial theorems*, preprint.

[19] A. J. HOFFMAN AND J. B. KRUSKAL, JR., *Integral boundary points of convex polyhedra*, in Linear Inequalities and Related Systems, Annals of Mathematics Studies 38, Princeton University Press, Princeton, NJ, 1956, pp. 223–246.

[20] A. J. HOFFMAN AND D. E. SCHWARTZ, *On partitions of a partially ordered set*, J. Combin. Theory (B), 23 (1977), pp. 3–13.

[21] G. O. H. KATONA, *On a conjecture of Erdös and a stronger form of Sperner's theorem*, Studia Sci. Math. Hungar., 1 (1966), pp. 59–63.

[22] ———, *A generalization of some generalizations of Sperner's theorem*, J. Combin. Theory, 12 (1972), pp. 72–81.

[23] ———, *A three part Sperner theorem*, Studia Sci. Math. Hungar., 8 (1973), pp. 379–390.

[24] D. J. KLEITMAN, *On a lemma of Littlewood and Offord on the distribution of certain sums*, Math. Z., 90 (1965), pp. 251–259.

[25] G. L. NEMHAUSER AND L. E. TROTTER, *Vertex packings: structural properties and algorithms*, Technical Report No. 210, Operations Research Dept., Cornell University, Ithaca, NY, January, 1974.

[26] G. W. PECK, *Maximum antichains of rectangular arrays*, J. Combin. Theory (A), 27 (1979), pp. 397–400.

[27] R. PROCTOR, M. SAKS AND D. G. STURTEVANT, *Product partial orders with the Sperner property*, preprint.

[28] M. SAKS, private communication.

[29] ———, *Duality properties of finite set systems*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge MA, 1980.

[30] ———, *A short proof of the existence of k-saturated partitions of a partially ordered set*, Advances Math., 33 (1979), pp. 207–211.

[31] J. SCHONHEIM, *A generalization of results of P. Erdös, G. Katona, and D. J. Kleitman concerning Sperner's theorem*, J. Combin. Theory (A), 11 (1971), pp. 111–117.

[32] E. SPERNER, *Ein Satz über Untermengen einer endlichen Mengë*, Math. Z., 27 (1928), pp. 544–549.

[33] D. B. WEST AND D. J. KLEITMAN, *Skew chain orders and sets of rectangles*, Discrete Math., 27 (1979), pp. 99–102.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 3, September 1981

0196-5212/81/0203-0010 $01.00/0

A HELLY THEOREM FOR SETS*

G. W. PECK†

Abstract. In this note we prove the following Helly type result.

THEOREM. *Let X be a collection of subsets of an n element set S with the property that any k members of X have an element in common. If X has at least $(k+2)2^{n-k-1}+1$ members, then all members of X have an element in common. The same statement fails for bounds of one less for $n \geqq k+1$.*

In this note we prove the following Helly type result.

THEOREM. *Let X be a collection of subsets of an n element set S with the property that any k members of X have an element in common. If X has at least $(k+2)2^{n-k-1}+1$ members, then all members of X have an element in common. The same statement fails for bounds of one less for $n \geqq k+1$.*

The argument used here is an example of application of the "pushing" method [1], [2] which can be described as follows. Imagine that our n elements are the integers 1 to n; then each set C containing j but not $j-1$ is "pushed" into the corresponding set with $j-1$ but not j by a "j-push"; no other sets are affected by a j-push. A collection of sets has each of its member sets pushed when a j-push is applied to it unless the resulting set is already in the collection, in which case it is left alone.

By a succession of such pushes for different values of j one can take a collection of sets into a "canonical form" having the same number of members which is invariant under all j-pushes. Statements of the form: "every k members of Z have intersection at least $j \ldots$" are preserved by pushing so that one can, by pushing, reduce discussion of collections restricted by only such properties to the possible push invariant collections.

To prove our theorem here, we apply this technique to a maximal sized collection Z of subsets of S such that every k or fewer have an element in common, but not all do. It is easy to see that the first condition of the previous sentence (that every k or fewer members have an element in common) is preserved by pushing. What is perhaps surprising is that the second condition, that not all elements of X have an element in common, is also preserved under pushing for maximal sized X.

The argument proceeds by proving the following observations:

1. If a set A is in X and B contains A, then B is in X.
2. All $(n-1)$ element sets lie in X.
3. Any push invariant Y obtained by pushing on X has every k members containing an element in common and not all its members with an element in common.
4. If no collection of k members of Y have only n in common, then the collection of subsets of $\{1, \cdots, n-1\}$ obtained by omitting n from all members of Y containing n obeys these same two conditions and, by induction, we have $|Y| = |X| \leqq 2(k+2)2^{n-1-k-1} = (k+2)2^{n-k-1}$.
5. If there are collections Q of k or fewer members of Y that contain only n in common, then complements of these elements must form a partition of $1, \cdots, n-1$.

* Received by the editors January 22, 1981. This research was supported in part by the Office of Naval Research under contract N00014-76-C-0366.

† Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

6. Every member of such a collection Q must be a minimal member of Y in that no subset of it can be in Y.
7. If any member of a Q has $(n-1)$ members, by induction $|Y| \leqq (k+1)2^{n-k-1}+1$.
8. Suppose that the hypothesis of 7 does not hold and that t members of Y lie in the union of all Q. Then one can find a set T containing $t/2$ or fewer of them that intersect every Q.
9. By removing the members of T from Y and adding the sets obtained by omitting n from all other members of the union of all Q, we obtain a collection $\bar{Y}$ of the same size or larger than Y with the same properties in which no collection of k members of $\bar{Y}$ have only n in common, and by 4 the result is proven.
10. There is a collection obeying the given properties having $(k+2)2^{n-k-1}$ elements for $n \geqq k+1$.

We now prove these observations.

1. If a collection of k or fewer members of $X \cup \{B\}$ has vanishing intersection, then the same collection with B replaced by A also does; therefore by the maximal size of X, B must be in X.

2. By virtue of 1, if an $(n-1)$ element set A is not in X, then all members of X contain the element complementary to A in S, violating the definition of X.

3. We show that the result of a j-push ($j(Y)$) of Y obeys the conditions given on X if Y did. First we show that it obeys the first property: that every k or fewer members have an element in common. If some set G of k or fewer members of Y has an element other than j in common, the images in $j(Y)$ do so as well. If G's members had only j in common, its image in $j(Y)$ would have j or $j-1$ in common unless there were a set G' in Y of the same size as G with empty intersection; G' can be obtained by taking the members of G with both j and $j-1$, those with j only that are not in $j(Y)$, and those with $j-1$ only whose presence in Y caused the others in G to be unaltered under the push. The $(n-1)$ element sets are obviously invariant under pushing, so that the second property holds as well.

5. If some element j of S was not in two of the members of Q, replacing n by j in one of them would yield a set which would be disjoint from the intersection of the remainder of Q.

6. If one replaced a member of Q by a proper subset, the argument of 5, here, would yield a collection of k or fewer members of Y with empty intersection.

7. If a member D of Q has $(n-1)$ elements, by 6 it must be the only member of Y without the element it lacks. The remaining members of Y must obey the conditions on Y for $(n-1)$ and $(k-1)$ since using D as one of the k members eliminates that element and reduces k to $(k-1)$ for the rest. Induction then yields $|Y| \leqq (k+1)2^{n-k-1}+1$.

8. For each Q we choose two members, the one not containing $n-1$, and any one other. We define a graph G among the resulting chosen set by connecting the two members chosen for each Q by an arc. We then seek here to find a set of vertices that intersect every arc, using at most half the vertices that appear in all the arcs. Since G is bipartite, it is well known that this can be done. (The number of vertices necessary to intersect all the arcs is the size of a maximum matching in the graph; this is a statement of Hall's [3] marriage theorem.)

9. The statement is a proof in itself.

10. For each set A or k or $k+1$ elements out of $1, \cdots, k+1$ take all sets of the form $A \cup B$ with $B = \{k+2, \cdots, n\}$. The resulting collection of sets obeys the given property and has $(k+2)2^{n-k-1}$ members, for $n \geqq k+1$.

For $n \leqq k$ any collection of sets such that any k of its members have an element in common necessarily has some element common to all its members.

Acknowledgment. The author thanks D. J. Kleitman for his help in writing this paper.

REFERENCES

[1] P. ERDÖS, CHAO KO AND R. RADO, *Intersection theorems for systems of finite sets*, Quart. J. Math. Oxford Sec (2), 12 (1961), pp. 313–318.

[2] D. J. KLEITMAN, *On a combinatorial conjecture of Erdös*, JCT, 1 (1966), pp. 209–214.

[3] P. HALL, *On representatives of subsets*, J. London Math. Soc., 10 (1935), pp. 26–30.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 3, September 1981

0196-5212/81/0203-0011 $01.00/0

A COUNTEREXAMPLE TO A BIN PACKING CONJECTURE*

JAMES B. SHEARER†

Abstract. In this note we exhibit a counterexample to the following conjecture of Garey, Graham and Johnson: If $L=\{a_1,\cdots,a_n\}$ is an ordered list of items with sizes $s(a_i)(0<s(a_i)\leqq 1)$ let $\mathrm{FF}(L)$ be the number of bins of size 1 required by the "first-fit" algorithm to pack L. Let $R(\alpha)=\limsup_{N\to\infty}[(1/N)\max\{\mathrm{FF}(L)|L$ can be packed into N bins of size $\alpha\}]$. Let $\omega(\alpha)=\max\{\sum_{i=1}^{k}1/(P_i-1)|2\leqq P_1\leqq P_2\leqq\cdots;\ \sum_{i=1}^{k}1/P_i=\alpha$, the P_i are integers at least 2 of which $\neq 2\}$. Then $R(\alpha)=\omega(\alpha)$.

Let $L=\{a_1, a_2,\cdots a_n\}$ be an ordered list of items along with a size function s which assigns to each $a_i\in L$ a size $s(a_i)$ satisfying $0<s(a_i)\leqq 1$. The first-fit (FF) bin packing algorithm packs the items of L into bins $B_1, B_2,\cdots$, each bin having a size associated with it, by successively placing each item of L in the bin of lowest index to which it will fit (a set of items fits into a bin if the sum of the item sizes does not exceed the bin size). Let $\mathrm{FF}(L)$ be the number of bins used by the FF algorithm when packing L into bins of size 1. Let $R(\alpha)=\limsup_{N\to\infty}[(1/N)\max\{\mathrm{FF}(L)|L$ can be packed into N bins of size $\alpha\}]$.

Suppose $\alpha=\sum_{i=1}^{k}1/P_i$ (k may be ∞), where the P_i are integers $\geqq 2$ at least two of which $\neq 2$. We assume $P_1\leqq P_2\leqq\cdots$. Then there exists a set of lists which show that $R(\alpha)\geqq\sum_{i=1}^{k}1/(P_i-1)$ (see [1]). Hence if we define $\omega(\alpha)=\max\{\sum_{i=1}^{k}1/(P_i-1)|2\leqq P_1\leqq P_2\leqq\cdots;\ \sum_{i=1}^{k}1/P_i\leqq\alpha$, the P_i are integers at least 2 of which $\neq 2\}$ we clearly have $R(\alpha)\geqq\omega(\alpha)$.

It is shown in [1] that the decomposition of α achieving $\omega(\alpha)$ is that one in which we successively choose $P_1, P_2\cdots$ to be as small as possible consistent with the conditions that the P_i are integers $\geqq 2$, at least two of which $\neq 2$ and $\sum_{i=1}^{k}1/P_i\leqq\alpha$. If α is rational this procedure will terminate after a finite number of terms. If α is irrational the procedure will generate an infinite number of terms but $\sum_{i=1}^{\infty}1/(P_i-1)$ will converge very rapidly.

For example, let $\alpha=1$. Then $P_1=2$, $P_2=3$ and $P_3=6$. We cannot choose $P_2=2$ because this would violate the condition that at least 2 of the $P_i\neq 2$. The procedure terminates after P_3 is chosen since $1-\frac{1}{2}-\frac{1}{3}-\frac{1}{6}=0$. Hence $\omega(1)=1+\frac{1}{2}+\frac{1}{5}=\frac{17}{10}$. It is known [3] that $R(1)=\frac{17}{10}$ also. This and other examples led Garey, Graham and Johnson to conjecture in [1] (see also [2]) that $R(\alpha)=\omega(\alpha)$ for all α. However the following example shows that this conjecture is false.

Let $\alpha=\frac{1}{3}+\frac{1}{7}+\frac{1}{62}=\frac{641}{1302}=\frac{2564}{5208}$. Let N be an arbitrary positive integer and consider the list L of $120N$ items with sizes

$$\begin{aligned}
s(a_i)&=\tfrac{745}{5208}, && 1\leqq i\leqq 30N,\\
s(a_i)&=\tfrac{869}{5208}, && 30N<i\leqq 60N,\\
s(a_{2i-1})&=\tfrac{1695}{5208}, && 30N<i\leqq 60N,\\
s(a_{2i})&=\tfrac{1819}{5208}, && 30N<i\leqq 60N.
\end{aligned}$$

L can be packed into $60N$ bins of α. In this packing $30N$ bins contain 1 item of size $\frac{745}{5208}$ and 1 item of size $\frac{1819}{5208}$ and the remaining $30N$ bins contain 1 item of size $\frac{869}{5208}$ and 1 item of size $\frac{1695}{5208}$.

* Received by the editors February 10, 1981.

† Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139. This work was done while the author was a consultant at Bell Laboratories, Murray Hill, New Jersey 07974.

The FF algorithm packs L into $41N$ bins of size 1. The first $5N$ bins each contain 6 items of size $\frac{745}{5208}$. The next $6N$ bins each contain 5 items of size $\frac{869}{5208}$. The remaining $30N$ bins each contain 1 item of size $\frac{1695}{5208}$ and 1 item of size $\frac{1819}{5208}$. Hence $R(\alpha) \geqq \frac{41}{60}$. But $\omega(\alpha) = \frac{1}{2} + \frac{1}{6} + \frac{1}{61} = \frac{2500}{3660} < \frac{2501}{3660} = \frac{41}{60}$. Hence $R(\alpha) \neq \omega(\alpha)$ for all α, so the conjecture is false.

This is not an isolated example. In fact, the constructions that show $R(\alpha) \geqq \omega(\alpha)$ can be modified to show $\lim_{\varepsilon \to 0+} R(\alpha - \varepsilon) \geqq \omega(\alpha)$. Since $\lim_{\varepsilon \to 0+} \omega(\alpha - \varepsilon) < \omega(\alpha)$ whenever α is rational numerous counterexamples exist. However, it can be shown that if in the expansion $\alpha = \sum_{i=1}^{k} 1/P_i$ achieving $\omega(\alpha)$ the P_i increase sufficiently fast then $R(\alpha) = \omega(\alpha)$. Determining $R(\alpha)$ for all α appears to be a difficult problem.

REFERENCES

[1] M. R. GAREY, R. L. GRAHAM AND D. S. JOHNSON, *On a number-theoretic bin packing conjecture*, Proc. 5th Hungarian Combinatorics Colloquium, North-Holland, Amsterdam, 1978, pp. 377–392.

[2] M. R. GAREY AND D. S. JOHNSON, *Approximation algorithms for bin packing problems: A survey*, preprint.

[3] 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. 2, No. 3, September 1981

0196-5212/81/0203-0012 $01.00/0

LOCAL PROPERTIES OF k-NN REGRESSION ESTIMATES*

Y. P. MACK†

Abstract. Let (X_i, Y_i), $i = 1, \cdots, n$ be i.i.d. bivariate random vectors such that $X_i \in \mathbb{R}^p$, $Y_i \in \mathbb{R}^1$. Suppose $r_n(x)$ denotes the k-nearest neighbor (k-NN) estimator of $r(x) = E(Y|X = x)$. Under appropriate conditions, we derive the rates of convergence for the bias and variance as well as asymptotic normality of $r_n(x)$. These appear to share some similarities with the k-NN density estimates. The technique is by conditioning on R_n, the Euclidean distance between x and its kth nearest neighbor among the X_i's. Some comparison is made between the k-NN and kernel methods.

1. Introduction. Let (X_i, Y_i), $i = 1, \cdots, n$ be independent identically distributed (i.i.d.) random vectors such that $X_i \in \mathbb{R}^p$, $Y_i \in \mathbb{R}^1$. Let $f(x, y)$ be the joint density of X and Y, $f(x) = \int f(x, y)\, dy$ be the marginal density of X and $r(x) = E(Y|X = x)$ be a version of the conditional expectation function defined via $f(x, y)$ and $f(x)$. The problem of estimating $r(x)$ from (X_i, Y_i) within the parametric setting has been well studied, in particular when $f(x, y)$ is assumed to be multivariate normal. In this discussion, we concern ourselves with a class of nonparametric estimates of $r(x)$. We shall investigate their asymptotic behavior and make some comparison with another class of nonparametric regression estimates.

Going through the research literature on the topic at hand (the reader can find a comprehensive listing in [14]), we realize that much less is known compared with nonparametric density estimation, although the two problems share some similarities. Among the sources, we mention the work of Watson [17] where he suggested the following estimators of $r(x)$ in the bivariate $f(x, y)$ case:

$$\tilde{r}_n(x) = \left[\frac{1}{n}\sum_{j=1}^{n} \delta_n(x - X_j)\right]^{-1} \cdot \left[\frac{1}{n}\sum_{j=1}^{n} \delta_n(x - X_j) \cdot Y_j\right], \tag{1}$$

where δ_n is a sequence of nonnegative weight functions tending to the Dirac delta function as $n \to \infty$, and

$$r_n(x) = \frac{1}{k}\sum_{j \in J} Y_j, \tag{2}$$

where $J = \{i: X_i$ is one of the $k = k(n)$ observations nearest to $x\}$. Watson gave some analysis of the bias and variance of (1) and remarked that perhaps (2) is easier to handle than (1) since it does not involve the ratio of two random quantities. (We shall return to this remark later.) (1) and (2) are noteworthy since they are precursors of two large classes of nonparametric regression estimates, namely, the kernel and the k-nearest neighbor (k-NN) methods. Along the line of (2), Royall [12] in his doctoral dissertation had made detailed analysis of the MSE, MISE,[1] as well as asymptotic normality of a generalization of (2) given by

$$r_n(x) = \sum_{i=1}^{n} c_{ni} W_{ni}, \tag{2}'$$

where $W_{ni} = Y_j$ if X_j is the ith closest observation to x, and c_{ni} defines a triangular array of nonnegative numerical weights.

* Received by the editors January 14, 1980, and in revised form November 24, 1980. This paper is based in part on the author's PhD dissertation at the University of California.

† Department of Mathematics, University of California at San Diego, La Jolla, California 92093. Currently at the Department of Statistics, University of Rochester, Rochester, New York, 14627.

[1] MISE in this case means $\int \text{MSE}\,(r_n(x))\, dF(x)$.

Proceeding as in density estimation, Rosenblatt [11] considered kernel type regression estimates which parallel (1):

$$(1)' \qquad \tilde{r}_n(x) = \left[\frac{1}{nb(n)} \sum_{j=1}^{n} w\left(\frac{x - X_j}{b(n)}\right)\right]^{-1} \cdot \left[\frac{1}{nb(n)} \sum_{j=1}^{n} w\left(\frac{x - X_j}{b(n)}\right) \cdot Y_j\right].$$

Here $w(u)$ is a weight function and $b(n)$ is a sequence of bandwidths satisfying $b(n) \to 0$, $nb(n) \to \infty$ as $n \to \infty$. Under appropriate regularity conditions on $f(x, y)$, Rosenblatt derived the bias, variance and asymptotic normality of (1)′ for $p = 1$, although it is clear that the higher dimensional case can be treated in the same manner.

Roughly speaking, since the density and regression functions are local in character, both the density and regression estimators can be regarded as appropriate averages of the observations in a neighborhood of the point under consideration. Whereas in the kernel method, where a deterministic region Σ based on $b(n)$ is formed about x, and then those sample points falling inside Σ are averaged; following an idea of Fix and Hodges [4] related to classification, the k-NN method first assigns a sequence of positive integers $k = k(n)$ with

$$(3) \qquad k \to \infty, \qquad \frac{k}{n} \to 0 \qquad \text{as } n \to \infty,$$

then the smallest sphere S containing the k nearest neighbors of x among the observations is located; finally an average of the k points is formed. In each method, the estimators are obtained after dividing the average by the volume of the appropriate region (Σ in the kernel case, S in the k-NN case). In this respect, we see that (1) and (2) are prototypes of the two methods if we let

$$\delta_n(u) = \frac{1}{2b(n)} 1_{\{|u| \leqq b(n)\}}$$

in (1), and write (2) as

$$r_n(x) = \left[\frac{1}{n} \sum_{j=1}^{n} \delta_n^*(x - X_j)\right]^{-1} \cdot \left[\frac{1}{n} \sum_{j=1}^{n} \delta_n^*(x - X_j) \cdot Y_j\right]$$

with

$$\delta_n^*(u) = \frac{1}{2R_n} 1_{\{|u| \leqq R_n\}}$$

where 1_A is the indicator on the set A and R_n is the distance between x and its kth nearest neighbor.

One ostensible advantage of the k-NN approach as suggested by a number of authors (see [2], [9], [16], for example) is that it is locally adaptive: if $f(x)$ is small, then S is large, and vice versa. Such a property is not enjoyed by the kernel method, as the volume of Σ remains the same for all x. The extent to which this property contributes to the local behavior of k-NN density estimates has been investigated in detail in Mack [7] and Mack and Rosenblatt [8], where a comparison with the kernel method was also made. On the other hand, the k-NN regression estimates, of which (2) and (2)′ are special cases, have been studied recently by Stone [13], [14] in a broader context and under rather mild assumptions. In particular, the k-NN regression estimates are L^2-consistent. Later Lai [6] in his thesis gave the MSE, MISE rates of convergence of

(2) by specifying

$$c_{ni} = \frac{1}{k} \quad \text{if } 1 \leqq i \leqq k,$$

$$= 0 \quad \text{otherwise}$$

in (2)′, in which case (for $p = 1$)

$$\text{Var}\,(r_n(x)) = \frac{\text{Var}\,(Y|X=x)}{k} + o\left(\frac{1}{k}\right), \tag{4}$$

and

$$\text{Bias}\,(r_n(x)) = \frac{1}{24f(x)^3}[(rf)''(x) - r(x)f''(x)] \cdot \left(\frac{k}{n}\right)^2 + o\left(\frac{k^2}{n^2}\right) + O\left(\frac{1}{k}\right).$$

An inspection of the bias expression reveals (as in the k-NN density estimates) that the scale factor involving $f(x)$ in the denominator can be nontrivially large in the tail region of $f(x)$. It would be of interest to see if such behavior persists in the higher dimensional case with a general weight function. In addition, as Bickel mentioned in a remark to Stone [14], (among other things), there is need to consider the asymptotic normality of these general k-NN estimates. This paper will attempt to answer these questions.

Throughout our discussion, we shall consider k-NN regression estimates given by

$$r_n(x) = \frac{h_n(x)}{f_n(x)}, \tag{5}$$

where

$$h_n(x) = \frac{1}{nR_n^p} \sum_{j=1}^{n} w\left(\frac{x - X_j}{R_n}\right) Y_j, \qquad f_n(x) = \frac{1}{nR_n^p} \sum_{j=1}^{n} w\left(\frac{x - X_j}{R_n}\right),$$

$w(u)$ is a bounded, nonnegative weight function satisfying

$$\int w(u)\, du = 1, \quad \text{and} \tag{6}$$

$$w(u) = 0 \quad \text{for } \|u\| \geqq 1. \tag{7}$$

R_n here will be defined according to the Euclidean norm $\|\cdot\|$ in $\mathbb{R}^p$, and $k(n)$ satisfies (3). (For a more detailed discussion of the role played by $\|\cdot\|$ on $\mathbb{R}^p$, where $\|\cdot\|$ need not be the Euclidean norm, see Stone [14].)

Note that (5) incorporates the features of both the kernel and the k-NN methods, and the weight w is a function of not only the distance but also the direction of the data with respect to x, whereas the numerical weights in (2)′ depend only on the nearest neighbor distance.

We first state the main results:

THEOREM 1. *Suppose f is bounded, $k = o(n)$, $\log n = o(k)$, w satisfies* (6), (7) *and*

$$\int \|v\|^2 |w(v)|\, dv < \infty, \qquad \int v_\alpha w(v)\, dv = 0, \qquad \alpha = 1, \cdots, p. \tag{8}$$

Suppose $P(\|x - X\| > r) = O(r^{-\zeta})$ for some $\zeta > 0$ as $r \to \infty$. Further suppose r and f are continuously differentiable up to second order in a neighborhood of x. Then if $f(x) > 0$, we

have

$$Er_n(x) = r(x) + \frac{[Q(rf)(x) - r(x)Q(f)(x)]}{2f(x)\cdot(cf(x))^{2/p}}\cdot\left(\frac{k}{n}\right)^{2/p} + o\left(\left(\frac{k}{n}\right)^{2/p}\right) + O\left(\frac{1}{k}\right). \tag{9}$$

Here

$$Q(g)(x) = \sum_{\alpha,\beta}\int v_\alpha v_\beta D_\alpha D_\beta g(x) w(v)\, dv,$$

and $c = \pi^{p/2}/\Gamma((p+2)/2) =$ *volume of unit ball in* $\mathbb{R}^p$.

THEOREM 2. *Suppose* $\int y^\beta f(x, y)\, dy$ *is bounded for* $\beta = 0, 1, 2$, $k = o(n)$, $\log n = o(k)$, *w satisfies* (6), (7) *and*

$$\int |v_\alpha|\cdot|w(v)|\, dv < \infty, \qquad \alpha = 1, \cdots, p. \tag{10}$$

Suppose $P(\|x - X\| > r) = O(r^{-\zeta})$ *for some* $\zeta > 0$ *as* $r \to \infty$. *Further suppose r and f are continuously differentiable in a neighborhood of x. Then if* $f(x) > 0$, *we have*

$$\operatorname{Var}(r_n(x)) = \frac{c\cdot \operatorname{Var}(Y|X = x)}{k}\int w^2(v)\, dv + o\left(\frac{1}{k}\right). \tag{11}$$

Remark. In view of Theorems 1 and 2, we have

COROLLARY. $r_n(x)$ *is pointwise consistent.*

THEOREM 3. *Suppose* $\int y^\beta f(x, y)\, dy$ *is bounded and continuous at x for* $\beta = 0, 1, 2$, *continuously differentiable in a neighborhood of x for* $\beta = 0, 1$. *Suppose* $E|Y|^3 < \infty$ *and w satisfies* (6), (7). *If* $k = o(n)$, $\log n = o(k)$, $\operatorname{Var}(Y|X = x) > 0$ *and* $f(x) > 0$. *Then*

$$\sqrt{k-1}\,[r_n(x) - Er_n(x)] \to N\left(0, c\cdot \operatorname{Var}(Y|X = x)\int w^2(v)\, dv\right) \tag{12}$$

in distribution as $n \to \infty$.

2. Preliminary remarks. As will be apparent, our technique is by conditioning on R_n. Denoting the common distribution of $\{\|x - X_j\|\}$ by G, we see that R_n is simply the kth order statistic from the i.i.d. sample $\{\|x - X_j\|\}$ with density

$$h(r) = n\binom{n-1}{k-1} G(r)^{k-1}(1 - G(r))^{n-k} G'(r), \tag{13}$$

where

$$\begin{aligned} G'(r) &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left\{\int_{\|t-x\|\leqq r+\varepsilon} f(t)\, dt - \int_{\|t-x\|\leqq r} f(t)\, dt\right\} \\ &= \int_{\|t-x\|=r} f(t)\, d\sigma(t), \end{aligned}$$

with σ denoting the surface area of $\|t - x\| = r$.

Under the assumption that $f(x)$ is continuous, almost surely, all the observations (X_i, Y_i) have distinct first coordinates. Let the $k-1$ observations with their first coordinates falling inside the sphere $\{z : \|z - x\| < r\}$ be denoted by $(\tilde{X}_i, \tilde{Y}_i)$, $i = 1, \cdots, k-1$. Then, conditioned on $R_n = r$, their joint density is given by

$$f(\tilde{x}_1, \cdots, \tilde{x}_{k-1}; \tilde{y}_1, \cdots, \tilde{y}_{k-1}|r) = \prod_{i=1}^{k-1}[f(\tilde{x}_i, \tilde{y}_i)/G(r)]. \tag{14}$$

Thus the $(\tilde{X}_i, \tilde{Y}_i)$ are conditionally independent, given $R_n = r$.

In order to have good enough estimates, we need to consider moments involving R_n in some detail. If we let

$$t = G(r) = P(\|x - X\| \leqq r)$$

$$= \int_{\|t-x\| \leqq r} f(t)\, dt \tag{15}$$

$$= f(x) \cdot \text{volume}\,\{z : \|z - x\| \leqq r\} + \int_{\|t-x\| \leqq r} [f(t) - f(x)]\, dt,$$

as $r \downarrow 0$,

$$t = f(x) \cdot cr^p + o(r^p),$$

where $c = \pi^{p/2}/\Gamma((p+2)/2) =$ volume of the unit ball in $\mathbb{R}^p$. Thus, as $r \downarrow 0$,

$$r = G^{-1}(t) = [cf(x)]^{-1/p} \cdot t^{1/p} + o(t^{1/p}). \tag{16}$$

Since $T = G(R_n)$ is just the kth order statistic from a uniform (0, 1) sample of size n, moments involving R_n can be computed via (16). The existence of such moments is ensured if we assume the tail-decay condition $P(\|x - X\| > r) = O(r^{-\zeta})$ for some $\zeta > 0$ as $r \to \infty$.

Next, suppose ϕ is a differentiable function; then by Taylor expansion we have, as $R_n \to 0$,

$$\phi(x - uR_n) = \phi(x - uG^{-1}(T))$$

$$= \phi(x - uG^{-1}(ET)) - \phi'(u - G^{-1}(ET) \cdot (G^{-1}(T) - G^{-1}(ET)) \cdot u$$

$$+ o(G^{-1}(T) - G^{-1}(ET))$$

$$= \phi\left(x - uG^{-1}\left(\frac{k}{n+1}\right)\right) - \phi'\left(u - G^{-1}\left(\frac{k}{n+1}\right)\right) \cdot u \cdot \left(\frac{1}{p}\right) \tag{17}$$

$$\cdot \left(\frac{1}{cf(x)}\right)^{1/p} \cdot \left(\frac{k}{n+1}\right)^{(1/p)-1} \left(T - \frac{k}{n+1}\right)$$

$$+ o\left(\frac{1}{p} \cdot \left(\frac{k}{n+1}\right)^{(1/p)-1} \left(T - \frac{k}{n+1}\right)\right).$$

Finally, for some real numbers a and b, we have

$$r_n(x) = \frac{a + h_n(x) - a}{b + f_n(x) - b}$$

$$= \frac{a}{b} + \frac{1}{b}(h_n(x) - a) - \frac{a}{b^2}(f_n(x) - b) + O[(f_n(x) - b)^2 + (h_n(x) - a)(f_n(x) - b)], \tag{18}$$

provided $|(1/b)f_n(x) - b)| < 1$ and $b \neq 0$.

In order that we can take expectation of the decomposition of $r_n(x)$ given in (18) (see Noda [10, § 4]), we develop the following result which has some interest in its own right.

LEMMA 1. *Suppose w is a positive bounded weight function satisfying* (6) *and* (7). *Suppose f is positive and continuous at x. If k satisfies $k/\log n \to \infty$, $k/n \to 0$ as $n \to \infty$ and $P(\|x - X\| > r) = O(r^{-\zeta})$ for some $\zeta > 0$ as $r \to \infty$, then $f_n(x) \to f(x)$ w.p.1 as $n \to \infty$.*

Proof. Write

$$f_n(x) = g_n(x) \cdot u_n(x), \tag{19}$$

where

$$g_n(x) = \frac{k-1}{cnR_n^p} \tag{20}$$

and

$$u_n(x) = \frac{c}{k-1} \sum_{j=1}^{k-1} w\left(\frac{x-\tilde{X}_j}{R_n}\right), \tag{21}$$

with the $\tilde{X}_j$'s defined earlier. Then the result of Devroye and Wagner [3] implies that $g_n(x) \to f(x)$ w.p.1 under our assumptions on k and f. It remains to show that

$$u_n(x) - E(u_n(x)|R_n) \to 0 \quad \text{w.p.1} \tag{22}$$

and that

$$E(u_n(x)|R_n) \to 1 \quad \text{w.p.1}. \tag{23}$$

Now using Theorem 2 of Hoeffding [5], by the conditional independence of the $\tilde{X}_j$'s

$$\begin{aligned} P\{u_n(x) - E(u_n(x)|R_n) > \varepsilon\} &= EP\{u_n(x) - E(u_n(x)|R_n) > \varepsilon | R_n\} \\ &\leqq \exp\left\{\frac{-2(k-1)\varepsilon^2}{N^2}\right\}, \end{aligned} \tag{24}$$

where N is the bound for w. Thus the Borel–Cantelli lemma implies (22) since $k \to \infty$ as $n \to \infty$. To show (23), note that

$$\begin{aligned} E(u_n(x)|R_n) &= \frac{c}{G(R_n)} \int w\left(\frac{x-t}{R_n}\right) f(t)\, dt \\ &= g_n(x)^{-1} \cdot \frac{k}{nG(R_n)} \cdot \int w(v) f(x - R_n v)\, dv, \end{aligned} \tag{25}$$

which tends to 1 w.p.1 under our assumptions.

3. Proofs of Theorems 1 and 2. We state a number of results from [8] as propositions here for later reference:

PROPOSITION 1. *Suppose f is bounded, w satisfies* (6), (7) *and* (8). *Suppose $P(\|x - X\| > r) = O(r^{-\zeta})$ for some $\zeta > 0$ as $r \to \infty$. Further suppose f is continuously differentiable up to second order in a neighborhood of x. Then if $f(x) > 0$, we have*

$$Ef_n(x) = f(x) + \frac{Q(f)(x)}{2(cf(x))^{2/p}} \left(\frac{k}{n}\right)^{2/p} + o\left(\left(\frac{k}{n}\right)^{2/p}\right). \tag{26}$$

PROPOSITION 2. *Suppose f is bounded, w satisfies* (6), (7) *and* (10). *Suppose $P(\|x - X\| > r) = O(r^{-\zeta})$ for some $\zeta > 0$ as $r \to \infty$. Further suppose f is continuously differentiable in a neighborhood of x. Then if $f(x) > 0$, we have*

$$\operatorname{Var}(f_n(x)) = \frac{cf^2(x)}{k} \int w^2(v)\, dv + o\left(\frac{1}{k}\right). \tag{27}$$

We now proceed to prove Theorems 1 and 2. For the remainder of this discussion, we assume that k satisfies (3) and that $k/\log n \to \infty$ as $n \to \infty$.

PROPOSITION 3. *Suppose* $\int yf(x, y)\,dy$ *is bounded and differentiable up to second order in a neighborhood of* x. *Suppose* w *satisfies* (6), (7) *and* (8). *Further suppose* $P(\|x-X\|>r)=O(r^{-\zeta})$ *for some* $\zeta>0$ *as* $r\to\infty$. *Then if* $f(x)>0$ *we have*

$$Eh_n(x)=r(x)f(x)+\frac{Q(rf)(x)}{2(cf(x))^{2/p}}\left(\frac{k}{n}\right)^{2/p}+o\left(\left(\frac{k}{n}\right)^{2/p}\right). \tag{28}$$

Proof.

$$\begin{aligned} E(h_n(x)|R_n)&=\frac{k-1}{nG(R_n)}\int w(v)\int yf(x-vR_n, y)\,dy\,dv \\ &=\frac{k-1}{nT}\int w(v)(rf)(x-vR_n)\,dv \qquad (29)\\ &=\frac{k-1}{nT}r(x)f(x)+\frac{k-1}{nT}\int w(v)[(rf)(x-vR_n)-(rf)(x)]\,dv. \end{aligned}$$

Since $(rf)(x)$ is differentiable up to second order in a neighborhood of x, and $\int v_\alpha w(v)\,dv=0$ for $\alpha=1,\cdots,p$, we have

$$\int w(v)[(rf)(x-vR_n)-(rf)(x)]\,dv=\tfrac{1}{2}Q(rf)(x)\cdot R_n^2+o(R_n^2). \tag{30}$$

Thus

$$\begin{aligned} E(h_n(x))&=E\left(\frac{k-1}{nT}\right)r(x)f(x)+\tfrac{1}{2}Q(rf)(x)\cdot E\left(\frac{k-1}{nT}\cdot R_n^2\right)+o\left(E\left(\frac{k-1}{nT}\cdot R_n^2\right)\right) \\ &=r(x)f(x)+\tfrac{1}{2}Q(rf)(x)\cdot\frac{1}{(cf(x))^{2/p}}\cdot\left(\frac{k}{n}\right)^{2/p}+o\left(\left(\frac{k}{n}\right)^{2/p}\right). \end{aligned} \tag{31}$$

PROPOSITION 4. *Suppose* $\int y^\beta f(x, y)\,dy$ *is bounded and continuous at* x *for* $\beta=1, 2$, w *satisfies* (6), (7) *and* (10). *Suppose* $\int yf(x, y)\,dy$ *is continuously differentiable in a neighborhood of* x, *and* $P(\|x-X\|>r)=O(r^{-\zeta})$ *for some* $\zeta>0$ *as* $r\to\infty$. *Let* $t(x)=E(Y^2|X=x)$, *then*

$$\operatorname{Var}(h_n(x))=\frac{cf^2(x)t(x)}{k}\int w^2(v)\,dv+o\left(\frac{1}{k}\right), \tag{32}$$

$$\operatorname{Cov}(h_n(x), f_n(x))=\frac{cr(x)f^2(x)}{k}\int w^2(v)\,dv+o\left(\frac{1}{k}\right). \tag{33}$$

Proof. First note that

$$\operatorname{Var}h_n(x)=E(\operatorname{Var}(h_n(x)|R_n))+\operatorname{Var}(E(h_n(x)|R_n)).$$

Now

$$\begin{aligned} \operatorname{Var}(h_n(x)|R_n)&=\frac{k-1}{n^2R_n^{2p}}\operatorname{Var}\left(w\left(\frac{x-X_j}{R_n}\right)\middle|R_n\right) \\ &=\frac{k-1}{n^2R_n^pT}\int w^2(v)(tf)(x-vR_n)\,dv-\frac{k-1}{nT^2}\left[\int w(v)(rf)(x-vR_n)\,dv\right]^2 \qquad (34)\\ &=①-②. \end{aligned}$$

Write

$$①=\frac{k-1}{n^2R_n^pT}\cdot(tf)(x)\int w^2(v)\,dv+\frac{k-1}{n^2R_n^pT}\int w^2(v)[(tf)(x-vR_n)-(tf)(x)]\,dv.$$

The expectation of the first term in ① is

$$\frac{cf^2(x)t(x)}{k}\int w^2(v)\,dv+o\left(\frac{1}{k}\right).$$

The expectation of the second term in ① can be shown to tend to zero by the Schwarz inequality and by the boundedness and continuity of $\int y^2 f(x, y)\,dy$ at x.

Similarly, we can show that if $\int yf(x, y)\,dy$ is bounded and continuous at x, then

$$E②=\frac{r^2(x)f^2(x)}{k}+o\left(\frac{1}{k}\right).$$

Thus we have

$$E(\mathrm{Var}\,(h_n(x)|R_n))=\frac{cf^2(x)t(x)}{k}\int w^2(v)\,dv-\frac{r^2(x)f^2(x)}{k}+o\left(\frac{1}{k}\right). \tag{35}$$

Next, from (29), we have

$$E(h_n(x)|R_n)=\frac{k-1}{nT}\int w(v)(rf)(x-vR_n)\,dv.$$

Using (17), we obtain

$$\begin{aligned}(rf)(x-vR_n)&=(rf)\left(x-vG^{-1}\left(\frac{k}{n+1}\right)\right)\\ &+\sum_\alpha v_\alpha D_\alpha(rf)\left(x-vG^{-1}\left(\frac{k}{n+1}\right)\right)\cdot\left(\frac{1}{p}\right)\\ &\cdot\left(\frac{1}{cf(x)}\right)^{1/p}\left(\frac{k}{n+1}\right)^{(1/p)-1}\cdot\left(T-\frac{k}{n+1}\right)\\ &+o\left(\frac{1}{p}\cdot\left(\frac{k}{n+1}\right)^{1/p-1}\cdot\left(T-\frac{k}{n+1}\right)\right).\end{aligned} \tag{36}$$

Proceeding as in the proof of [8, Prop. 2], for $|T-k/(n+1)|\geqq k/(n+1)$, the contribution to $\mathrm{Var}\,(E(h_n(x)|R_n))$ in this range can be shown to be $O(e^{-k/2})$. For $|T-k/(n+1)|<k/(n+1)$, we have

$$\frac{k-1}{nT}=1-\left(\frac{k}{n+1}\right)^{-1}\left(T-\frac{k}{n+1}\right)+O\left(\left(\frac{k}{n+1}\right)^{-1}\cdot\left(T-\frac{k}{n+1}\right)^2\right), \tag{37}$$

so that in order to estimate the variance of (29), we only need to consider the second moment of

$$\begin{aligned}&-\int w(v)(rf)\left(x-vG^{-1}\left(\frac{k}{n+1}\right)\right)\cdot\left(\frac{k}{n+1}\right)^{-1}\left(T-\frac{k}{n+1}\right)dv\\ &-\left[\sum_\alpha\int v_\alpha D_\alpha\left(x-vG^{-1}\left(\frac{k}{n+1}\right)\right)w(v)\,dv\right]\cdot\left(\frac{1}{p}\right)\\ &\cdot\left(\frac{1}{cf(x)}\right)^{1/p}\cdot\left(\frac{k}{n+1}\right)^{(1/p)-1}\cdot\left(T-\frac{k}{n+1}\right)\\ &+o\left(\frac{1}{p}\left(\frac{k}{n+1}\right)^{(1/p)-1}\cdot\left(T-\frac{k}{n+1}\right)\right).\end{aligned}$$

Since Var $T = k/n^2$ as $n \to \infty$, the second moment can be shown to be equal to

$$\frac{r^2(x)f^2(x)}{k} + o\left(\frac{1}{k}\right).$$

To summarize, we have

$$\text{Var}\,(E(h_n(x)|R_n)) = \frac{cf^2(x)t(x)}{k}\int w^2(v)\,dv + o\left(\frac{1}{k}\right)$$

and (32) is proved.

In a similar manner, it can be shown that

$$\text{Cov}\,(h_n(x), f_n(x)) = \frac{cr(x)f^2(x)}{k}\int w^2(v)\,dv + o\left(\frac{1}{k}\right).$$

Theorem 1 now follows from (18), (26), (27), (28), (33), Lemma 1, and Theorem 2 follows from (18), (27), (32), (33) and Lemma 1.

Remark. Comparison between the kernel and the k-NN regression estimates.

We consider the case where $p=1$ with $w(t) = \frac{1}{2} \cdot 1_{\{|t|<1\}}$ in some detail. The results shown in Table 1 on the kernel estimate are extracted from Rosenblatt [11].

TABLE 1

	kernel	k-NN
bias	$\frac{[(rf)''(x) - r(x)f''(x)]}{2f(x)} \cdot \int v^2 w(v)\,dv \cdot b^2(n) + o(b^2(n)) + O((nb(n))^{-1})$	$\frac{[(rf)''(x) - r(x)f''(x)]}{8f^3(x)} \cdot \int v^2 w(v)\,dv \left(\frac{k}{n}\right)^2 + o\left(\left(\frac{k}{n}\right)^2\right) + O\left(\frac{1}{k}\right)$
variance	$\frac{\text{Var}\,(Y\|X=x)}{f(x)\cdot nb(n)}\int w^2(v)\,dv + o\left(\frac{1}{nb(n)}\right)$	$\frac{2\,\text{Var}\,(Y\|X=x)}{k}\int w^2(v)\,d + o\left(\frac{1}{k}\right)$
Optimal rate	Optimal rate	
Mean-squared error	$b(n) = O(n^{-1/5})$ $\text{MSE} = O(n^{-4/5})$	$k(n) = O(n^{4/5})$ $\text{MSE} = O(n^{-4/5})$

The correspondence $nb(n) \leftrightarrow k(n)$ seems to hold as in density estimation. Here, the scale factors in the bias and variance terms seem to indicate that in low density regions the bias in the k-NN estimate can be quite large; while the variance may do better.

In the general case, the MSE of the kernel estimate is minimized by setting $b(n) = dn^{-\lambda}$, $0 < \lambda < 1$, and then we have $\lambda = 1(p+4)$ and the optimal rate of decay of the MSE is $O(n^{-4/(4+p)})$. The constant d in this case is given by

$$d^{p+4} = p \cdot \frac{\text{Var}\,(Y|X=x)}{f(x)} \cdot \frac{f^2(x)\int w^2(v)\,dv}{[Q(rf)(x) - r(x)Q(f)(x)]^2[\int v^2 w(v)\,dv]^2} \tag{38}$$

and

$$\text{MSE} = \left[\frac{\text{Var}\,(Y|X=x)}{f(x)}\int w^2(v)\,dv\right]^{4/(4+p)}\left[\frac{1}{2f(x)}(Q(rf)(x)-r(x)Q(f)(x))\right]^{2p/(4+p)}$$

$$\cdot\left[\left(\frac{4}{p}\right)^{p/(4+p)}+\left(\frac{p}{4}\right)^{4/(p+4)}\right]\cdot n^{-4/(4+p)}. \tag{39}$$

For the k-NN estimate, if we set $k(n)=d'n^{\gamma}$, $0<\gamma<1$, the optimal rate is obtained for $\gamma=4/(4+p)$, whence the rate of decay for the MSE is $O(n^{-4/(4+p)})$ also. The constant d' is given by

$$(d')^{(p+4)/p}=\left(\frac{p}{4}\right)\left[\frac{c\ \text{Var}\,(Y|X=x)4(cf(x))^{4/p}\cdot f^2(x)\int w^2(v)\,dv}{[Q(rf)(x)-r(x)Q(f)(x)]^2\cdot[\int v^2 w(v)\,dv]^2}\right], \tag{40}$$

and the MSE is found to be the same as (39) again.

4. Proof of Theorem 3. Let us write

$$\psi_n(x)=\sqrt{k-1}\,[r_n(x)-Er_n(x)]=A_n(x)+B_n(x)\,, \tag{41}$$

where

$$A_n(x)=\sqrt{k-1}\,[r_n(x)-E(r_n(x)|R_n)],$$
$$B_n(x)=\sqrt{k-1}\,[E(r_n(x)|R_n)-Er_n(x)].$$

We shall show that $A_n(x)\to N(0, c\ \text{Var}\,(Y|X=x)\int w^2(v)\,dv)$ in distribution and $B_n(x)\to 0$ in probability as $n\to\infty$. First, we have the following.

LEMMA 2. (a) *Suppose f is bounded and continuous at x and w satisfies* (6), (7). *Then*

$$E(f_n(x)|R_n)\to f(x) \quad \textit{in probability as } n\to\infty, \tag{42}$$

$$k\cdot\text{Var}\,(f_n(x)|R_n)\to\left[cf^2(x)\int w^2(v)\,dv-f^2(x)\right] \quad \textit{in probability as } n\to\infty. \tag{43}$$

(b) *Suppose $\int yf(x, y)\,dy$ is bounded and continuous at x and w satisfies* (6), (7). *Then*

$$E(h_n(x)|R_n)\to r(x)f(x) \quad \textit{in probability as } n\to\infty, \tag{44}$$

$$k\cdot\text{Cov}\,(h_n(x), f_n(x)|R_n)\to\left[cf^2(x)r(x)\int w^2(v)\,dv-r(x)f^2(x)\right] \quad \textit{in probability as } n\to\infty. \tag{45}$$

(c) *Suppose $\int y^{\beta}f(x, y)\,dy$ is bounded and continuous at x for $\beta=1, 2$, and w satisfies* (6), (7). *Then*

$$k\cdot\text{Var}\,(h_n(x)|R_n)\to\left[cf^2(x)t(x)\int w^2(v)\,dv-r^2(x)f^2(x)\right] \quad \textit{in probability as } n\to\infty. \tag{46}$$

Proof. The results follow from the fact that

$$\frac{k-1}{nR_n^p}\to cf(x) \quad \text{in probability}$$

and

$$\frac{k-1}{nT} \to 1 \quad \text{in probability}$$

as $n \to \infty$.

PROPOSITION 5. *Suppose* $\int y^\beta f(x, y)\, dy$ *is bounded and continuous at* x *for* $\beta = 0, 1, 2$, w *satisfies* (6), (7) *and* $E|Y|^3 < \infty$. *Assume* $\operatorname{Var}(Y|X = x) > 0$. *Then if* $f(x) > 0$

$$A_n(x) \to N\Big(0, c \operatorname{Var}(Y|X = x) \int w^2(v)\, dv\Big)$$

in distribution as $n \to \infty$.

Proof. Let $a_n = E(h_n(x)|R_n)$, $b_n = E(f_n(x)|R_n)$. Then by Lemma 2 and (18), as $n \to \infty$, we can write

$$r_n(x) = \frac{a_n}{b_n} + \frac{1}{b_n}(h_n(x) - a_n) - \frac{a_n}{b_n^2}(f_n(x) - b_n) + o_p(1). \tag{47}$$

Conditioned on R_n, let $(\tilde{X}_i, \tilde{Y}_i)$ be the $k-1$ observations such that $\|x - \tilde{X}_i\| < R_n$. Let

$$Z_j = \frac{k-1}{nR_n^p}\left[\frac{\tilde{Y}_j}{b_n} - \frac{a_n}{b_n^2}\right] \cdot w\left(\frac{x - \tilde{X}_j}{R_n}\right). \tag{48}$$

Then, conditioned on R_n,

$$A_n(x) = [\operatorname{Var}(Z_j|R_n)]^{1/2} \cdot \phi_n(x), \tag{49}$$

where

$$\phi_n(x) = \frac{1}{\sqrt{k-1}} \sum_{j=1}^{k-1} [Z_j - E(Z_j|R_n)]/[\operatorname{Var}(Z_j|R_n)]^{1/2}$$

is a centered and normalized average of i.i.d. random variables. As in Lemma 2, it can be shown that

$$\operatorname{Var}(Z_j|R_n) \to c \operatorname{Var}(Y|X = x) \int w^2(v)\, dv > 0 \tag{50}$$

in probability as $n \to \infty$.

Under the assumption that $E|Y|^3 < M < \infty$ and $\sup w < N < \infty$, if we denote the conditional distribution of $\phi_n(x)$, given $R_n = r$, by $F_n(t|r)$, and the distribution of the standard normal by $\Phi(t)$, then the Berry–Esseen central limit theorem implies that, for some constant $\theta > 0$,

$$|F_n(t|r) - \Phi(t)| \leqq \frac{\theta \cdot MN}{\sqrt{k-1}\,[\operatorname{Var}(Z_j|R_n)]^{3/2}} \tag{51}$$

as $n \to \infty$.

For $\varepsilon > 0$ small enough, recalling the density of R_n from (13), we have

$$\begin{aligned} &|P(\phi_n(x) \leqq t) - \Phi(t)| \\ &\quad \leqq \int |F_n(t|r) - \Phi(t)| h(r)\, dr \\ &\quad \leqq \frac{\theta MN}{\varepsilon^{3/2}\sqrt{k-1}} P(\operatorname{Var}(Z_j|R_n) > \varepsilon) + P(\operatorname{Var}(Z_j|R_n) \leqq \varepsilon) \to 0 \quad \text{as } n \to \infty \end{aligned} \tag{52}$$

by (50). This completes the proof.

PROPOSITION 6. *Suppose* $\int y^{\beta} f(x, y)\, dy$ *is bounded and continuously differentiable at* x *for* $\beta = 0, 1$, w *satisfies* (6), (7). *If* $f(x) > 0$, *then* $B_n(x) \to 0$ *in probability as* $n \to \infty$.

Proof. Expression (29) and

$$E(f_n(x)|R_n) = \frac{k-1}{nT} f(x) + \frac{k-1}{nT} \int w(v)[f(x - vR_n) - f(x)]\, dv \tag{53}$$

imply that

$$E(r_n(x)|R_n) = \left\{ r(x) + \frac{1}{f(x)} \int w(v)[(rf)(x - vR_n) - (rf)(x)]\, dv - \frac{r(x)}{f(x)} \int w(v)[f(x - vR_n) - f(x)]\, dv \right\} \cdot (1 + o(1)). \tag{54}$$

Using (36) and

$$f(x - vR_n) = f\left(x - vG^{-1}\left(\frac{k}{m+1}\right)\right) - \sum_{\alpha} v_{\alpha} D_{\alpha} f\left(x - vG^{-1}\left(\frac{k}{n+1}\right)\right) \cdot \frac{1}{p} \cdot \left(\frac{1}{cf(x)}\right)^{1/p} \cdot \left(\frac{k}{n+1}\right)^{(1/p)-1} \cdot \left(T - \frac{k}{n+1}\right) \cdot (1 + o(1)), \tag{55}$$

we find that the contribution to Var $(\sqrt{k-1}\ (r_r(x)|R_n))$ comes from the second moment of

$$\left\{ \sum_{\alpha} \int v_{\alpha} \left[\frac{D_{\alpha}(rf)}{f(x)} - \frac{r(x) D_{\alpha}(f)}{f(x)} \right] \left(x - vG^{-1}\left(\frac{k}{n+1}\right) \right) w(v)\, dv \right\} \cdot \frac{1}{p} \left(\frac{1}{cf(x)}\right)^{1/p} \cdot (\sqrt{k-1}) \cdot \left(\frac{k}{n+1}\right)^{(1/p)-1} \cdot \left(T - \frac{k}{n+1}\right) \cdot (1 + o(1)). \tag{56}$$

But this is $O((k/n)^{2/p})$, since Var $T = k/n^2$ as $n \to \infty$. Thus Chebychev's inequality implies $B_n(x) \to 0$ in probability as $n \to \infty$.

Theorem 3 is now a consequence of Lemma 2 and Propositions 5 and 6.

5. Other comments. Despite the drawback in the bias behavior of the k-NN regression estimates in the low density region, the variances appear to do better than the kernel estimates. Also the MSE's of both types of estimates for optimal choices of the bandwidth $b(n)$ and the sequence $k(n)$ turn out to be identical (39). One advantage of the k-NN regression estimates which is not shared by the kernel estimates is that they are invariant under a scale change on the X-variable, as pointed out in [12]. Next, returning to the remark of Watson as indicated in the Introduction, for $p = 1$ with uniform weights one can exploit the properties of order statistics and their corresponding concomitants. If we define the r.v.'s $W_j = \|x - X_j\|$, and define the concomitant $Y_{[j]}$ by

$$Y_{[j]} = Y_i \quad \text{if } W_{(j)} = W_i.$$

where $W_{(j)}$ is the jth order statistic from the i.i.d. sample $\{W_j\}$, then $R_n = W_{(k)}$, and we can write (2) as

$$r_n(x) = \frac{1}{k} \sum_{j=1}^{k} Y_{[j]},$$

i.e., $r_n(x)$ can be studied as a linear combination of concomitants. Results on bias, variance and asymptotic normality can be obtained by appealing to a paper of Yang [18]. Recently, some optimal rate of convergence results of $r_n(x)$ have been obtained by Stone [15]. Finally, some indications of the performance of the kernel and k-NN regression estimates with computer simulated data are illustrated in the works of Benedetti [1] and Stone [13].

Acknowledgment. This paper is based in part on the author's 1978 Ph.D. dissertation completed at the University of California, San Diego, under the supervision of Professor Murray Rosenblatt. The author would also like to thank the referee for the constructive comments.

REFERENCES

[1] J. K. BENEDETTI, *Kernel estimation of regression functions*, in Proc. of Computer Science and Statistics, 8th Annual Symp. of the Interface, Health Science Computing Facility, UCLA, Los Angeles, 1975, pp. 405–412.

[2] L. BRIEMAN, W. MEISEL AND E. PURCELL, *Variable kernel estimates of multivariate densities and their calibration*, Technometrics, 19 (1977), pp. 135–144.

[3] L. P. DEVROYE AND T. J. WAGNER. *The strong uniform consistency of nearest neighbor density estimates.* Ann. Statist. 5 (1977), pp. 536–540.

[4] E. FIX AND J. L. HODGES, *Discriminatory analysis; nonparametric discrimination: consistency properties*, USAF SAM Series in Statistics, Project No. 21-49-004, Report No. 4, 1951.

[5] W. HOEFFDING, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc., 58 (1963), pp. 13–30.

[6] S. L. LAI, *Large sample properties of k-nearest neighbor procedures*, Ph.D. dissertation, University of California, Los Angeles, 1977.

[7] Y. P. MACK, *Asymptotic normality of multivariate k-NN density estimates*, Sankhya, Ser. A, (1981), to appear.

[8] Y. P. MACK AND M. ROSENBLATT, *Multivariate k-nearest neighbor density estimates.* J. Mult. Analysis, 9 (1979), pp. 1–15.

[9] D. S. MOORE AND J. W. YACKEL, *Large sample properties of nearest neighbor density function estimators*, in Statistical Decision Theory and Related Topics, Academic Press, New York, 1976.

[10] K. NODA, *Estimation of a regression function by the Parzen kernel-type density estimators*, Ann. Inst. Statist. Math., 28, (1976), pp. 221–234.

[11] M. ROSENBLATT *Conditional probability density and regression estimates*, in Multivariate Analysis II, Krishnaiah, ed., 1969, pp. 25–31.

[12] R. M. ROYALL, *A class of nonparameteric estimates of a smooth regression function*, Ph.D. dissertation, Stanford University, Stanford CA, 1966.

[13] C. J. STONE, *Nearest neighbor estimators of a nonlinear regression function*, in Proc. of Computer Science and Statistics: 8th Annual Symp. on the Interface, Health Science Computing Facility, UCLA, Los Angeles, 1975, pp. 413–418.

[14] ———, *Consistent nonparametric regression.* Ann. Statist., 5 (1977), pp. 595–620.

[15] ———, *Optimal rates of convergence for nonparametric inference.* Technical Report, UCLA, 1978.

[16] T. J. WAGNER, *Nonparametric estimates of probability density.* IEEE Trans. Inform. Theory, IT-21 (1975), pp. 438–440.

[17] G. S. WATSON, *Smooth regression analysis.* Sankhyā Ser. A, 26 (1964), pp. 359–372.

[18] S. S. YANG, *Linear functions of concomitants of order statistics*, Technical Report 7, Department of Mathematics, M.I.T., Cambridge MA, 1977.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 3, September 1981

0196-5212/81/0203-0013 $01.00/0

THE SCHENSTED CORRESPONDENCE AND LEXICOGRAPHIC MATCHINGS ON MULTI-SUBSET LATTICES*

KIEM-PHONG VO†

Abstract. We present a connection between lexicographic matchings on multisubset lattices, and the well-known Schensted correspondence. The connection is made via a natural representation of multisubsets as biwords. Some other interesting properties of the matchings are also given.

1. Introduction. Let $\bar{m}$ be the set $\{1, 2, \cdots, m\}$. A multiset is a set in which elements can be repeated. A multisubset of $\bar{m}$ is a multiset with elements from $\bar{m}$. Let $I = (i_1, i_2, \cdots, i_m)$ be a positive integral m-vector. A multisubset M of $\bar{m}$ is said to have I repetitions if the element k of M is repeated at most i_k times. Let S_m^I be the set of all multisubsets of $\bar{m}$ with I repetitions. It is easy to see that S_m^I is a distributive lattice under set inclusion. In the special case when all i_k's equal 1, S_m^I reduces to the usual subset lattice S_m of $\bar{m}$. A multisubset of $\bar{m}$ can always be represented by a weakly increasing sequence. Using this representation, the elements in each rank row of S_m^I can be totally ordered lexicographically.

It has been known for a long time that S_m has chain decompositions in which a chain starting at level i will end at level $m-i$. That is, S_m has symmetric matchings. In 1972, M. Aigner [1] explicitly constructed a symmetric matching for S_m that is compatible with the lexicographic ordering of rank rows. Informally, this is done as follows. A chain is started with the lexicographic least unused element in the lowest row. Each element in the chain is matched to the lexicographic least unused element of the row above that covers it if this element exists. The new element is marked used. The process is repeated until it cannot be further applied. Now a new chain is started. A matching constructed this way is called a lexicographic matching. Different constructions of this matching were discussed by other authors (for example, de Bruijn, Tengbergen, and Kruyswijk [2], Greene and Kleitman [4]). In 1977, D. E. White and S. G. Williamson [7] characterized all different constructions of this matching based on well-known recursions of $|S_m|$.

In this paper, we make a natural connection between lexicographic matchings on the lattice S_m^I and the well-known Schensted correspondence, relating biwords and pairs of tableaux of the same shape. In the process, it will become clear that the above matching on S_m is only a special case of the one we constructed on S_m^I.

2. The codings of multisubsets. As already mentioned above, a possible representation of multisubsets is by using weakly increasing sequences. We shall discuss two more representations, one by biwords, and the other by pairs of tableaux. The next few paragraphs will be devoted to clarifying these concepts.

Let A, B be two totally ordered finite sets (also called alphabets). A *biletter* is an element of $A \times B$ written as a column. We say that $\binom{a_1}{b_2} < \binom{a_2}{b_2}$ iff either $a_1 < a_2$ or $a_1 = a_2$ and $b_1 > b_2$. It is clear that $<$ defines a total order on $A \times B$. This is called the *locally reverse lexicographic order.* Let M be any multiset of biletters; we represent M by the sequence of biletters obtained by ordering elements of M in $<$. Such a sequence of biletters has a weakly increasing top row, and, in each block of equal letters in the top row, the corresponding letters in the second row are weakly decreasing (hence, "locally

* Received by the editors January 6, 1981.
† Department of Mathematics, University of California, San Diego, La Jolla, California 92093.

reverse"). A sequence of biletters so ordered is called a *biword*. We shall in general omit parentheses in a biword.

A *partition* $\lambda \vdash n$ of a positive number n is a sequence of positive integers $(\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_k)$ whose sum is n. The *Ferrers diagram* of λ is a collection of n cells arranged with λ_1 cells in the top row, λ_2 cells in the second row, etc. A *tableau* in A with shape λ is a filling of the Ferrers diagram λ with elements of A so that every row is weakly increasing and every column is strictly increasing. Here we are using the convention that rows are read from left to right, and columns are read from bottom to top.

Example 1. Let $A = \{1, 2, 3 \cdots n\}$, and $B = \{a, b, c \cdots\}$. Then:

(i) $\begin{matrix} 11122333 \\ bbacacba \end{matrix}$ is a biword in $A \times B$.

(ii) $\begin{matrix} 44 \\ 233 \\ 1122 \end{matrix}$ is a tableau in A with shape (2, 3, 4).

The *column insertion* is a process, found by Schensted [6], to insert an element a into a given tableau T in such a way that the tableau conditions are preserved. The process can be represented as follows:

```
procedure C–INSERT(a, T)
begin
  if T = ∅ then T := a;
  else begin
    Let c_1 < c_2 ··· < c_k be the first column of T,
      and T' be the rest of T;
    if a > c_k then append a to the first column;
    else begin
      Let i be so that c_{i-1} < a ≦ c_i;
      t := c_i;
      c_i := a;
      C–INSERT(t, T');
    end
  end
end
```

Let $\xrightarrow{c}$ denote column insertion. If $\alpha = a_1 a_2 \cdots a_k$ is a sequence of elements of A, we can obtain $T(\alpha)$, the tableau of α as:

$$T(\alpha) = a_k \xrightarrow{c} (\cdots (a_2 \xrightarrow{c} (a_1 \xrightarrow{c} \varnothing)) \cdots)$$

More interesting, given a biword $\left(\begin{smallmatrix} a_1 & a_2 \cdots a_k \\ b_1 & b_2 \cdots b_k \end{smallmatrix}\right)$, we can obtain a unique pair of tableaux of the same shape (T_B, T_A) as

```
procedure C-ENCODE( a_1 a_2 ··· a_k , T_B, T_A )
                  ( b_1 b_2 ··· b_k )
begin
  if k > 0 then
  begin
    C-ENCODE( a_1 a_2 ··· a_{k-1} , T_B, T_A );
            ( b_1 b_2 ··· b_{k-1} )
    C-INSERT(B_K, T_B);
    Put a_k at the corresponding new cell in T_A;
  end
end
```

We now have a theorem due to Schensted, Knuth and Burge:

THEOREM 1. *C*–ENCODE(∗, ϕ, ϕ) *defines a bijection between the set of biwords and the set of pairs of tableaux* (T_B, T_A) *of the same shape.*

These algorithms and their variations are discussed in [3], [5], [6].

We are now ready to give the other two representations of elements of S.

Let $B = \{0, 1\}$. Let

$$a = \underbrace{1 \cdots 1}_{k_1}\underbrace{2 \cdots 2}_{k_2} \cdots \underbrace{m \cdots m}_{k_m}$$

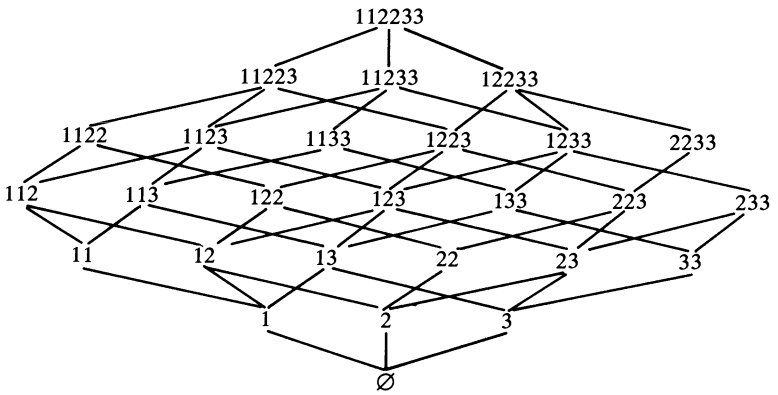
be an element in S_m^I. We code a with the following biword of $\bar{m} \times B$:

$$\begin{array}{c} \overbrace{1 \cdots 1}^{i_1}\overbrace{2 \cdots 2}^{i_2} \cdots \overbrace{m \cdots m}^{i_m} \\ \underbrace{1 \cdots 1}_{k_1}0 \cdots 0\underbrace{1 \cdots 1}_{k_2}0 \cdots 0 \cdots \underbrace{1 \cdots 1}_{k_m}0 \cdots 0 \end{array}$$

By the above theorem, each $a \in S_m^I$ corresponds uniquely to a pair of tableaux (P, Q). P, Q have the same shape with at most two rows. The entries of P are 0, 1's, while Q contains exactly i_1 1's, i_2 2's, etc. Generally, if T has i_1 1's, i_2 2's, $\cdots$ we say T is an I-tableau. In the special case when all i_k's equal 1, T is also called a standard tableau. Thus, the Q tableaux are I-tableaux.

So we have three different codings of an element in S_m^I, by weakly increasing sequences, biwords, or pairs of tableaux. We shall not always explicitly mention which representation of an element is being used.

Example 2. Let $\bar{m} = \bar{3} = \{1, 2, 3\}$, and $I = (2, 2, 2)$; then $S_3^{(2,2,2)}$ is:



Also:

a) 1233	↔	112233	↔	11	23
		101011		0011,	1123
b) 123	↔	112233	↔	11	23
		101010		0001,	1123
c) 23	↔	112233	↔	11	23
		001010		0000,	1123

The fact that these elements of $S_3^{(2,2,2)}$ have the same right tableaux is not accidental. As we shall see next, they belong to the same chain in a lexicographic matching of $S_3^{(2,2,2)}$.

3. The matchings. Generally, as we have seen, if $a \in S_m^I$, then a can be represented uniquely as a pair of tableaux of the form

$$(*) \qquad \begin{array}{l} \overbrace{1\cdots 1}^{s_1'} \\ 0\cdots 0\underbrace{0\cdots 0}_{s_0}\underbrace{1\cdots 1}_{s_1}, Q \end{array}$$

in which any of the strings s_0, s_1, s_1' can be empty, and Q is an I-tableau. When $s_0 \neq \varnothing$ we can define:

DEFINITION.

$$\mathrm{NEXT}(a) = \begin{array}{l} 1\cdots 1 \\ 0\cdots 0\underbrace{0\cdots 0}_{s_0'}\underbrace{1\cdots 1}_{s_1}, Q \end{array}$$

That is, NEXT(a) is the pair of tableaux (P', Q) obtained by switching the last 0 of the string s_0 to 1.

It is clear from the definition, and Theorem 1 that NEXT(a) is in S_m^I. So NEXT is an injection that lifts elements from a rank row of S_m^I to elements of the next rank. If $a \in S_m^I$ with representation (*), we let:

$$a_s = \begin{array}{l} 1\cdots 1 \\ 0\ \cdots\ 0, Q \end{array} \qquad \text{and} \qquad a_f = \begin{array}{l} 1\cdots 1 \\ 0\cdots 01\cdots 1, Q. \end{array}$$

Again, by Theorem 1, $a_s \in S_m^I$, and $a_f \in S_m^I$. Then NEXT defines a chain $C = a_s$, $\mathrm{NEXT}(a_s), \cdots, \mathrm{NEXT}^k(a_s) = a_f$. In the chain C, all elements have the same right tableau Q. We say that Q is the right tableau of C. We now have:

THEOREM 2. NEXT(a) *covers a in* S_m^I. *Thus,* NEXT *decomposes* S_m^I *into disjoint chains such that if a chain C starts at level i then*:

a) *The shape of tableaux of C is* $(i, N-i)$, *where* $N = \sum_{k=1}^{m} i_k$.

b) *The start and end elements of C have left tableaux*:

$$\begin{array}{l} \overbrace{1\cdots 1}^{i} \\ \underbrace{0\ \cdots\ 0}_{N-i} \end{array} \qquad \textit{and} \qquad \begin{array}{l} \overbrace{1\cdots 1}^{i} \\ \underbrace{0\cdots 01\cdots 1}_{N-i} \end{array}$$

Proof. We induct on m. The case $m = 1$ is trivial and omitted. Now, assume the assertions for $m-1$, and consider m. First, consider the case $i_m = 1$. In this case, we can partition $S_m^I = S_{m-1}^I \cup S_{m-1}^I(m)$, where S_{m-1}^I is the restriction of S_m^I in $\overline{m-1}$, and $S_{m-1}^I(m)$ is obtained by adding m to each element of S_{m-1}^I. By the induction hypothesis, all assertions are satisfied in S_{m-1}^I. We now examine what happens to the tableau representations of elements of S_{m-1}^I when S_{m-1}^I is embedded in S_m^I and $S_{m-1}^I(m)$. Let

$$\begin{array}{l} w = 1\ \cdot\ \cdot\ \cdot\ 12\ \cdot\ \cdot\ \cdot\ 2\cdots(m-1)\cdots(m-1) \\ \quad 1\cdots 10\cdots 01\cdots 10\cdots 0\cdots\ 1\cdots 10\cdots 0 \end{array}$$

be in S_{m-1}^I. There are 3 cases:

Case (i). w is considered as an element in S_m^I. The effect on the biword is that of adding the biletter $\binom{m}{0}$ to w. So

$$w \leftrightarrow \begin{array}{ll} 1\cdots 1 & q_1\cdots q_k \\ 0\cdots 00\cdots 01\cdots 1, & q_1'\ \cdots\ q_{N-k-1}' \end{array}$$

gives

$$w\binom{m}{0} \leftrightarrow \begin{matrix} 1\cdots 1 & & q_1\cdots q_k \\ 0\cdots 000\cdots 01\cdots 1, & q_1' & \cdots \quad q'_{N-k-1}m \end{matrix}$$

Now an element in $S^I_{m-1}(m)$ is one in S^I_{m-1} with the biletter $\binom{m}{1}$ appended. We have:

Case (ii). Let $w \in S^I_{m-1}$ be an end element of a chain. We have

$$\begin{matrix} w \leftrightarrow 1\cdots 1 & & q_1\cdots q_k \\ 0\cdots 01 \quad \cdots \quad 1, & q_1' & \cdots \quad q'_{N-k-1} \end{matrix}$$

so:

$$w\binom{m}{1} \leftrightarrow \begin{matrix} 1\cdots 1 & & q_1\cdots q_k \\ 0\cdots 011 \quad \cdots \quad 1, & q_1' & \cdots \quad q'_{N-k-1}m \end{matrix}$$

Case (iii). If w is not an end element of a chain in S^I_{m-1}

$$\begin{matrix} w \leftrightarrow 1\cdots 1 & & q_1\cdots q_k \\ 0\cdots 00\cdots 01\cdots 1, & q_1' & \cdots \quad q'_{N-k-1} \end{matrix}$$

and

$$w\binom{m}{1} \leftrightarrow \begin{matrix} 1\cdots 11 & & q_1\cdots q_k m \\ 0\cdots 00\cdots 01\cdots 1, & q_1'\cdots q_k' & \cdots q'_{N-k-1}, \end{matrix}$$

From (i), (ii) we see that end elements of chains in S^I_{m-1} are mapped to those considered in (ii), so that the covering property is preserved. These elements become new end elements of the chains involved. By (iii) the covering property is verified in $S^I_{m-1}(m)$. Note that the start elements of the new chains are those of S^I_{m-1} and those of $S^I_{m-1}(m)$ not considered in (ii). Properties (a), (b) follow easily from this observation. So, we are done with the case $i_m = 1$.

In general, if $i_m = k > 1$, we can partition:

$$S^I_m = S^I_{m-1} \cup S^I_{m-1}(m) \cup \cdots \cup S^I_{m-1}\underbrace{(m\cdots m)}_{k \text{ times}}.$$

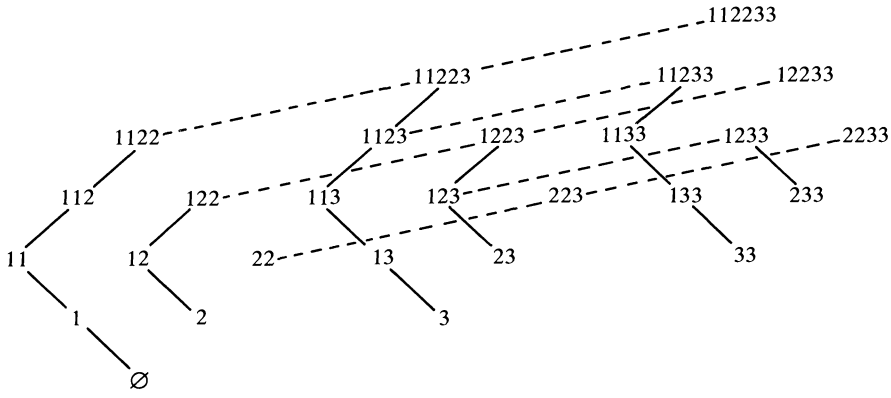
Now, we can induct on k. The verifications here are similar to the above and will be left as an exercise. □

An immediate corollary to the theorem is:

THEOREM 3. NEXT *decomposes* S^I_m *into disjoint symmetric chains. Every chain is associated uniquely with an I-tableau with at most two rows* (*the right tableau of the chain*). *Further, every such tableau is the right tableau of a chain defined by* NEXT.

Proof. The first two assertions follow from conditions (a), (b) of Theorem 2. Now, let Q be any I-tableau with at most two rows. Let P be the tableau with the same shape as Q with 1's on the top row, 0's in the second row. Theorem 1 shows that (P, Q) is some element of S^I_m. This element must be in some chain C. Then Q is the right tableau of C.

Example 3. The construction for $S_3^{(2,2,2)} = S_2^{(2,2,2)} \cup S_2^{(2m2m2)}(3) \cup S_2^{(2,2,2)}(33)$:



The chains of $S_3^{(2,2,2)}$ and their right tableaux:

∅	2	3	22	23	33	233
1	12	13	223	123	133	
11	122	113	2233	1233	1133	
112	1223	1123				
11223	12233	11233				
112233						
	2	3	22	23	33	233
112233	11233	11223	1133	1123	1122	112

By examining the proof of Theorem 2 more carefully (cases (ii) and (iii)) or by direct considerations of the tableau representation, we have

COROLLARY 4. *Let C be a chain defined by* NEXT *with right tableau Q. Then*

(a) *The top row elements of Q constitute the multiset that is the start element of C.*

(b) *Using the row deletion algorithm (see* [5], [7]*), sequentially delete corners in the top row of Q. The remaining elements constitute the end element of C.*

Example 4. Let C be the chain (23, 123, 1233) in $S_3^{(2,2,2)}$. Then the Q tableau of C is $\begin{smallmatrix} 23 \\ 1123 \end{smallmatrix}$. Following is the sequence of top row corner deletions:

$$\begin{matrix} 23 & & 2 \\ 1123 & \rightarrow & 1133 \end{matrix} \rightarrow 1233$$

So far, we have shown that NEXT defines a symmetric matching on S_m^I. It remains to show that NEXT is also lexicographic. For this, we need the following specialization of Theorem 3 in the case $I = (1, 1, \cdots 1)$.

LEMMA 5. NEXT *decomposes the subset lattice S_m into symmetric chains. Every chain is associated uniquely with a standard tableau with at most two rows. Moreover, every such tableau is the right tableau for some chain defined by* NEXT.

Proof. Let N be as in Theorem 2. Replace the top row of each biword (element) of S_m^I by the sequence $123 \cdots N$. Using the simple fact that this operation commutes with C–ENCODE, or a direct proof using similar arguments as in the proof of Theorem 2, we see that every chain defined by NEXT in S_m^I is associated with a unique standard tableau with N entries and at most two rows.

Example 5. The right tableaux of $S_3^{(2,2,2)}$ and corresponding standard tableaux are

$$\begin{array}{llll} 112233 & \leftrightarrow 123456, & & \\ 2 & \leftrightarrow 3 & 3 & \leftrightarrow 5 \\ 11233 & 12456, & 11223 & 12346, \\ 22 & \leftrightarrow 34 & 23 & \leftrightarrow 35 \\ 1133 & 1256, & 1123 & 1246, \\ 33 & \leftrightarrow 56 & 233 & \leftrightarrow 356 \\ 1122 & 1234, & 112 & 124. \end{array}$$

The new biwords obtained by replacing the top rows by $12 \cdots N$ can be viewed as the images of a natural embedding of S_m^I into S_N. So, we can consider S_m^I as a sublattice of S_N. From the above observation, we see that the chains of S_m^I form a subset of the set of chains in S_N. Now, on S_n, it is easy to see that NEXT is the same as the map τ defined by D. E. White and S. G. Williamson [7]. So, by their results, NEXT defines a lexicographic matching on S_N. Further, our embedding of S_m^I into S_N preserves the lexicographic ordering of elements. So, we have:

THEOREM 6. NEXT *defines a lexicographic matching on* S_m^I.

4. Further results. It is a famous theorem of Dilworth that the maximum size of an antichain is the same as the minimum number of chains needed in a chain partition for any poset. This number is called the Dilworth number. Let $N' = \sum_{k=1}^{m-1} i_k$, and $I' = (i_1, i_2, \cdots, i_{m-1}, N')$. Let $\mathscr{C}(S_m^I)$ be the set of chains defined by NEXT in S_m^I. Let $d(S_m^I)$ be the Dilworth number of S_m^I. From Dilworth's theorem, it is easy to see that $d(S_m^I) = |\mathscr{C}(S_m^I)|$. From the one-to-one correspondence between chains in S_m^I and I-tableaux, we have the following easy but somewhat surprising result:

THEOREM 7. *If* $i_m \geqq N'$, *then* $d(S_m^I) = d(S_m^{I'})$.

Proof. Since tableaux have strictly increasing columns, the number of possible top rows for I-tableaux with at most two rows is the same as that for I'-tableaux.

The chains of S_m^I can be partially ordered in a nice way. Let C_1, C_2 be chains of S_m^I; we say $C_1 \leqq C_2$ iff every element of C_1 is contained in the interval of S_m^I defined by the start and end elements of C_2. Clearly, by the symmetry of chains, and the fact that they partition S_m^I, $(\mathscr{C}(S_m^I), \leqq)$ is a partially ordered set, which we shall call the lexicographic chain poset of S_m^I. We have:

LEMMA 8. *Let* $C \in \mathscr{C}(S_m^I)$, *and* $w \in S_m^I$. *Let* a_s *and* a_f *be the start and end elements of* C *respectively. Let* $C(w)$ *be the unique element in* $\mathscr{C}(S_m^I)$ *containing* w. *If* $a_s \leqq w \leqq a_f$, *then* $C(w) \leqq C$.

Proof. Again, we employ the basic partition of S_m^I used in Theorem 2, and induct on m. We only consider the case $i_m = 1$. The reader can make appropriate extension to the general case following the techniques used in Theorem 2. The case $m = 1$, being trivial, will be omitted. Now, for general m, we partition $S_m^I = S_{m-1}^I \cup S_{m-1}^I(m)$ as in Theorem 2. The chains in S_m^I divide into two types, type A with start element in S_{m-1}^I, and type B with start element in $S_{m-1}^I(m)$. There are 4 cases:

Case (i). $w \in S_{m-1}^I$, C is type B: impossible because $a_s \not\leqq w$.

Case (ii). $w \in S_{m-1}^I$, C is type A: since w does not contain any m, if a_f^- is the element immediately preceding a_f in C, we have $w \leqq a_f^-$ (as $a_f^- = a_f - \{m\}$). By induction, the chain $C'(w)$ in S_{m-1}^I is contained in $[a_s, a_f^-]$ (where $[\cdot, \cdot]$ denotes an interval in S_m^I). Let w' be the end element of $C'(w)$; then $w' \cup \{m\}$ is the end element of $C(w)$ by construction. Thus, $C(w) \leqq [a_s, a_f]$.

Case (iii). $w \in S_{m-1}^I(m)$, C is type A. We have two subcases:

(a). $C(w)$ is type A. Then w must be the end element of $C(w)$. Let w^- be the

element immediately preceding w in $C(w)$. We must have $a_s \leqq w^- \leqq a_f^-$ where a_f^- is defined as above. Now, by induction, the chain $C(w^-)$ in S^I_{m-1} is contained in $[a_s, a_f^-]$. This again leads to $C(w) \leqq [a_s, a_f]$.

(b). $C(w)$ is type B. Remove all m from elements of $C(w)$. The remaining multisets form a chain $C'(w)$ of S^I_{m-1} without an end element. Again, by induction, $C'(w) \leqq [a_s, a_f]$, implying $C(w) \leqq [a_s, a_f]$.

Case (iv). $w \in S^I_{m-1}(m)$, C is type B. All involved elements contain an m. Remove this m, then repeat the basic induction argument, we have $C(w) \leqq [a_s, a_f]$.

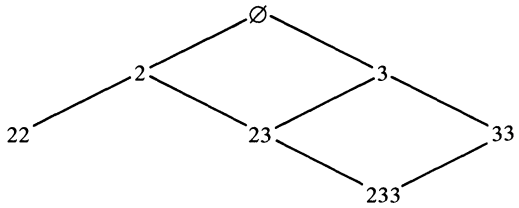
In any poset $(P, \leqq)$, let $x \in P$. Then the order ideal generated by x, $I(x)$ is the set $\{p \leftarrow P : p \leqq x\}$. We now have:

THEOREM 9. *Let $A \in \mathscr{E}(S^I_m)$, and let a_s, a_f be its start and end elements, respectively. Then $[a_s, a_f] = \bigcup_{C \in I(A)} C$.*

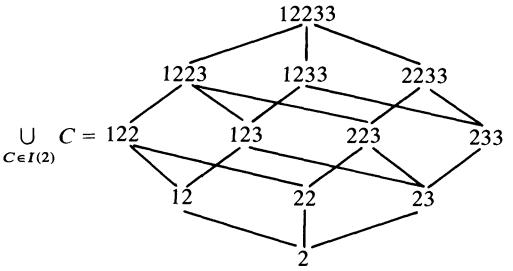
Proof. Certainly, we have $\bigcup_{C \in I(A)} C \leqq [a_s, a_f]$. Now let w be in $[a_s, a_f]$. Let $C(w)$ be the unique chain of S^I_m containing w. By Lemma 8, $C(w) \in I(A)$. So $[a_s, a_f] \leqq \bigcup_{C \in I(A)} C$.

From Theorem 3 and Corollary 4, every element in $\mathscr{E}(S^I_m)$ is uniquely identified by the top row of its associated right tableau. So, we shall label each chain with the top row of its right tableau in the following example.

Example 6. The chain poset of $(\mathscr{E}(S_3^{(2,2,2)}), \leqq)$:



$I(2) = \{2, 22, 23, 233\}$, and:



We end this section with a few numerical identities that are direct consequences of Theorems 2 and 3. Let s_k be the number of I-tableaux with shape $(k, N-k)$ $(k \leqq N/2)$, where N is defined as in Theorem 2. Let r_k be the number of elements with rank k in S^I_m. For $k \leqq N/2$, we have $s_k = r_k - r_{k-1}$. Let t_k be the number of tableaux with shape $(k, N-k)$ and 0, 1-entries. It can be seen directly that $t_k = N - 2k + 1$. We have:

THEOREM 10.

(a) $|S^I_m| = \sum_{k=0}^{\lfloor N/2 \rfloor} s_k t_k = \sum_{k=0}^{\lfloor N/2 \rfloor} s_k (N - 2k + 1)$,

(b) $r_{\lfloor N/2 \rfloor} = \sum_{k=0}^{\lfloor N/2 \rfloor} s_k$;

when all i_k's equal 1:

(c) $|S_m| = 2^m = \sum_{k=0}^{\lfloor m/2 \rfloor} \binom{m}{k}(m-2k+1)/(m-k+1).$

(d) $r_{\lfloor m/2 \rfloor} = \binom{m}{\lfloor m/2 \rfloor} = \sum_{k=0}^{\lfloor m/2 \rfloor} \binom{m}{k}(m-2k+1)/(m-k+1).$

Proof. (a), (b) follow from Theorems 2, 3. For (c), (d) we note that for $I = (1, \cdots 1)$, we have $N = m$, and:

$$s_k = r_k - r_{k-1} = \binom{m}{k} - \binom{m}{k-1} = \binom{m}{k}(m-2k+1)/(m-k+1).$$

Notes and acknowledgments. Corollary 4(a) and parts (a), (b) of Theorem 10 were also proved by C. Greene, and D. Kleitman.

The idea of using the Schensted correspondence in connection with chain matchings resulted from conversations between Profs. A. M. Garsia and D. E. White. I am grateful to Prof. Garsia for showing me this idea, also for his valuable encouragement, and advices. I would like to thank Roger Whitney for several important suggestions concerning this work, especially his assistance in the proof of Theorem 2.

REFERENCES

[1] M. AIGNER, *Lexicographic matching in Boolean algebras*, J. Combin. Theory Ser. B, 14 (1973), pp. 187–194.

[2] N. DE BRUIJN et al. *On the set of divisors of a number*, Nieuw Arch. Wisk., 23 (1951), pp. 191–193.

[3] W. H. BURGE, *Four correspondences between graphs and generalized Young tableaux*, J. Combin. Theory Ser. A, 17 (1974), pp. 12–30.

[4] C. GREENE AND D. KLEITMAN, *Strong versions of Sperner's theorem*, J. Combin. Theory Ser. A, 20 (1976), pp. 80–88.

[5] D. E. KNUTH, *Permutations, matrices and generalized Young tableaux*, Pacific J. Math., 34 (1970), pp. 709–727.

[6] C. SCHENSTED, *Longest increasing and decreasing subsequences*, Canad. J. Math., 13 (1961), pp. 179–191.

[7] D. E. WHITE AND S. G. WILLIAMSON, *Recursive matching algorithms and linear orders on the subset lattices*, J. Combin. Theory Ser. A, 23 (1977), pp. 117–127.

[8] K. P. VO AND R. WHITNEY, *Tableaux and matrix correspondences*, Univ. California, San Diego, 1980, preprint.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 3, September 1981

0196-5212/81/0203-0014 $01.00/0

INCIDENCE MATRICES OF SUBSETS—A RANK FORMULA*

NATHAN LINIAL† AND BRUCE L. ROTHSCHILD‡

Abstract. Let $n \geqq k \geqq l \geqq 0$ be integers, $\mathbb{F}$ a field, and $X = \{1, \cdots, n\}$. $M = M_{n,l,k}$ is an $\binom{n}{l} \times \binom{n}{k}$ matrix whose rows correspond to l-subsets of X, and columns to k-subsets of X. For $L \in X^{(l)}$, $K \in X^{(k)}$ the (L, K) entry of M is 1 if $L \subset K$, 0 otherwise. The problem is to find the rank of M over the field $\mathbb{F}$. We solve the problem for $\mathbb{F} = \mathbb{Z}_2$ and obtain some result on $\mathbb{F} = \mathbb{Z}_3$. The problem originated in extremal set theory and seems to be applicable also for matroids, codes and designs.

Introduction. The following problem was posed by M. Katchalski and M. A. Perles. Given $n \geqq k \geqq l \geqq 0$, integers, let $X = \{1, 2, \cdots, n\}$. Denote by $X^{(k)}$ the family of all subsets of X of cardinality k. A family of k-sets $\mathcal{K} \subset X^{(k)}$ is said to be closed if, for every $L \in X^{(l)}$, $|\{K \in \mathcal{K} | L \subset K\}|$ is never 1. They wanted to know the smallest number $N = N(n, l, k)$ such that if $\mathcal{A} \subset X^{(k)}$ has more than N sets, then it contains a closed subfamily. For $k = l+1$, their problem was solved by P. Frankl, who showed that in this case $N = \binom{n-1}{l-1}$. In fact he showed that if $\mathcal{A} \subset X^{(l+1)}$, has more than $\binom{n-1}{l-1}$ sets, then there is a family $\mathcal{K} \subset \mathcal{A}$, such that for every $L \in X^{(l)}$, $|\{K \in \mathcal{K} | L \subset K\}|$ is even. Define a matrix M whose rows (columns) are indexed by $X^{(l)}$ (resp. $X^{(l+1)}$). For $L \in X^{(l)}$, $K \in X^{(l+1)}$, the (L, K) entry is 1 if $L \subset K$, 0 otherwise. Frankl's proof is obtained by showing that the rank of this matrix over $\mathbb{Z}_2$ is $\binom{n-1}{l}$.

This raises the general problem: Given $n \geqq k \geqq l \geqq 0$, integers and a field $\mathbb{F}$, define a matrix $M = M_{n,l,k}$ as follows. Let $X = \{1, \cdots, n\}$, then the rows (columns) of M are indexed by $X^{(l)}$ (resp. $X^{(k)}$). For $L \in X^{(l)}$, $K \in X^{(k)}$, the (L, K) entry of M is 1 if $L \subset K$, 0 otherwise. What is the rank of M over the field $\mathbb{F}$? For $\mathbb{F} = \mathbb{Q}$ the answer appears in the literature [1], [2]; it is $\rho(M) = \min\{\binom{n}{l}, \binom{n}{l}\}$, so M has the highest rank possible. In this paper we solve the problem for $\mathbb{F} = \mathbb{Z}_2$ and for $k = l+1$ over $\mathbb{Z}_3$.

Define a *cycle* to be a family of k-sets such that every l-set is contained in an even number of these k-sets (this is usually done in algebraic topology). The rank formula over $\mathbb{Z}_2$ gives the largest cardinality of a cycle-free subfamily of $X^{(k)}$.

The rank formula over $\mathbb{Z}_2$. Let s be a nonnegative integer; we define $b(s)$ to be the unique set of nonnegative integers S, for which $s = \sum_{x \in S} 2^x$. Of course, b is an injective function. If p, q are integers with $b(p) \supset b(q)$ we simply write $p \supset q$. This defines a partial ordering on the nonnegative integers.

Define $d = k - l$, and let $D = b(d)$. For a function $f: D \to \mathbb{Z}^+$, the nonnegative integers we define $f(D) = \sum_{x \in D} f(x)$.

THEOREM 1. *For $n \geqq k + l$ the rank of $M_{n,l,k}$ over $\mathbb{Z}_2$ is*

$$\sum_{f: D \to \mathbb{Z}^+} (-1)^{f(D)} \binom{n}{l - \sum_{x \in D} f(x) 2^x}.$$

Notation. We denote the matrix $M_{n-p,l-q,k-r}$ by $[p, q, r]$, where p, q, r are nonnegative integers. Also, $[p, q, r]_l$ stands for $M_{n-p,l-q,l-r}$, and $[p, q, r]_k =$

* Received by the editors May 20, 1980, and in revised form February 12, 1981.

† Department of Mathematics, University of California, Los Angeles, CA 90024. The work of this author was supported by a Haim Weizman Postdoctoral Fellowship.

‡ Department of Mathematics, University of California, Los Angeles, CA 90024. The work of this author was partially supported by NSF Grant MCS79-037-11. The authors wish to thank the authorities of the grants for their kind support.

$M_{n-p,k-q,k-r}\langle p, q\rangle$ is defined to be the sum

$$\sum_{f:D\to Z^+}(-1)^{f(D)}\binom{n-p}{l-q-\sum_{x\in D}f(x)2^x}.$$

Observe that $M_{n,l,k}$ and $M_{n,n-k,n-l}$ are transposed matrices. Therefore, to cover the case $n \leqq l+k$ in Theorem 1, replace l by $n-k$ in the sum formula.

We need some simple observations which we state without proof.

Observation 1.

$$[0,0,0]=\begin{array}{|c|c|}\hline [1,1,1] & 0\\ \hline [1,0,1] & [1,0,0]\\ \hline\end{array}.$$

where the left (right) columns correspond to k-subsets which contain the element 1, (do not contain 1, resp.). The upper (lower) rows are the l-sets containing (not containing) 1.

Observation 2. For $p \leqq q \leqq r$, $M_{n,p,q}\cdot M_{n,q,r}=M_{n,p,r}\cdot\binom{r-p}{r-q}$.

Observation 3. $\binom{a}{b}$ is odd iff $a \supset b$.

Observation 4. $\langle p, q\rangle=\langle p+1, q\rangle+\langle p+1, q+1\rangle$.

Convention. If A is a matrix which depends on n, l, k, then $A(p, q, r)$ denotes the matrix which is obtained by replacing n by $n-p$, l by $l-q$ and k by $k-r$. Similarly, if A depends only on n and l (n and k), then $A(p, q)$ results on replacing n by $n-p$ and l by $l-q$ ($k-q$, resp.).

Let t be a nonnegative integer; then we define

$$S_t=\sum_{j\subset t}\langle t, j\rangle.$$

Also we define a block matrix A_t, indexed by all j such that $j \subset t$. Let $b(t)=\{a_1, \cdots, a_\tau\}$ with $a_1>a_2\cdots>a_\tau \geqq 0$. For $i, j \subset t$ the (i, j) block of A_t is $[t, i, j]$ if $j \supset i$ and $b(j-i)=\{a_1, \cdots, a_\nu\}$ for some $\nu \geqq 0$. All the other blocks are zero. Note that

$$S_0=\langle 0,0\rangle, \qquad A_0=[0,0,0],$$

and so we want to show that $\rho(A_0)=S_0$. Defining α by $2^\alpha \| d$, we prove the stronger:

PROPOSITION 1. *For* $0 \leqq t \leqq 2^\alpha$, $\rho(A_t)=S_t$.

Proof. By induction on n. For $n=0, 1$ there is nothing to prove. To perform the inductive step, we show that under the induction hypothesis the following hold:

PROPOSITION 2. $\rho(A_{2^\alpha})=S_{2^\alpha}$.

PROPOSITION 3. *For* $0 \leqq t \leqq 2^\alpha$, $\rho(A_{t+1})=S_{t+1}$ *implies* $\rho(A_t)=S_t$.

It is clear how Proposition 1 follows from Propositions 2, 3 by a backward induction.

Proof of Proposition 2. For $t=2^\alpha$, $b(t)=\{\alpha\}$, so:

$$A_t=\begin{array}{|c|c|}\hline [t,t,t] & \\ \hline [t,0,t] & [t,0,0]\\ \hline\end{array}, \qquad S_t=\langle t,0\rangle+\langle t,t\rangle.$$

The matrices

$$I+\begin{array}{|c|c|}\hline 0 & [t,t,0]_l \\ \hline 0 & 0 \\ \hline \end{array}, \qquad I+\begin{array}{|c|c|}\hline 0 & [t,t,0]_k \\ \hline 0 & 0 \\ \hline \end{array}$$

are nonsingular (in fact they are self-inverse), and they satisfy

$$\left(I+\begin{array}{|c|c|}\hline & [t,t,0]_l \\ \hline & \\ \hline \end{array}\right)A_t\left(I+\begin{array}{|c|c|}\hline & [t,t,0]_k \\ \hline & \\ \hline \end{array}\right)=\begin{array}{|c|c|}\hline & \\ \hline [t,0,t] & \\ \hline \end{array}.$$

To prove this, use Observations 2, 3 to show that in $\mathbb{Z}_2$ $[t,t,0]_l[t,0,0]=[t,t,0]\cdot\binom{d+t}{t}=0$, since $t=2^\alpha\|d$, and so $(d+t)\not\supset t$. Similarly $[t,t,t][t,t,0]_k=0$. But $[t,t,0]_l[t,0,t]=[t,t,t]\binom{d}{t}=[t,t,t]$, since $d\supset t$, and also $[t,0,t][t,t,0]_k=[t,0,0]$ for the same reason.

Rank is preserved under multiplying by the nonsingular matrices, and so $\rho(A_t)=\rho([t,0,t])$. From the induction hypothesis the last rank is

$$\sum_{f:D\setminus\{\alpha\}\to\mathbb{Z}^+}(-1)^{f(D\setminus\{\alpha\})}\binom{n-t}{l-\sum_{x\in D\setminus\{\alpha\}}f(x)2^x}.$$

Now $S_t=\langle t,0\rangle+\langle t,t\rangle=\sum_{f:D\to\mathbb{Z}^+}(-1)^{f(D)}\left[\binom{n-t}{l-\sum_{x\in D}f(x)2^x}+\binom{n-t}{l-t-\sum_{x\in D}f(x)2^x}\right]$.

All the second summands appear also as first summands with the opposite sign: increase $f(\alpha)$ by one. Doing all the canceling, we obtain only the sum of the first terms in which $f(\alpha)=0$; i.e.,

$$\sum_{f:D\setminus\{\alpha\}\to\mathbb{Z}^+}(-1)^{f(D\setminus\{\alpha\})}\binom{n-t}{l-\sum_{x\in D\setminus\{\alpha\}}f(x)2^x},$$

$\rho(A_t)=S_t$ for $t=2^\alpha$. □

Now we turn to the proof of Proposition 3. We establish a relation between A_t and A_{t+1}, between S_t and S_{t+1}. We define λ by $2^\lambda\|(t+1)$.

PROPOSITION 4.

$$S_t=S_{t+1}+2\sum_{\substack{j\subset t\\ 2^\lambda\not| j}}\langle t+1,j\rangle.$$

PROPOSITION 5.

$$\rho(A_t)=\rho(A_{t+1})+2\sum_{0\leqq\nu<\lambda}\rho(A_{t+1-2^{\nu+1}}(2^{\nu+1},2^\nu,2^\nu)).$$

First we show how Propositions 4, 5 imply Proposition 3. For any $0 \leqq \nu < \lambda$, set $r = t+1-2^{\nu+1}$. Using the inductive hypothesis we use the equality $\rho(A_r) = S_r = \sum_{j \subset r} \langle r, j \rangle$ with n replaced by $n-2^{\nu+1}$, l by $l-2^{\nu}$ and k by $k-2^{\nu}$; i.e. we use

$$\rho(A_r(2^{\nu+1}, 2^{\nu}, 2^{\nu})) = \sum_{j \subset r} \langle r+2^{\nu+1}, j+2^{\nu} \rangle = \sum_{j \subset r} \langle t+1, j+2^{\nu} \rangle = \sum_{\substack{i \subset t \\ 2^{\nu} \| i}} \langle t+1, i \rangle.$$

The last equality follows on setting $i = j+2^{\nu}$. Summing over all $0 \leqq \nu < \lambda$ yields that $\rho(A_{t+1}) = S_{t+1}$ implies $\rho(A_t) = S_t$; i.e., Proposition 4, 5 imply Proposition 3 and thus the main theorem.

We make the following simple observation.

Observation 5. For two nonnegative integers a, b, $a \subset b+1$ holds iff exactly one of the relations $a \subset b$, $a-1 \subset b$ holds.

Proof of Proposition 4. $S_t = \sum_{j \subset t} \langle t, j \rangle$, and by Observation 4 it equals $\sum_{j \subset t} \langle t+1, j \rangle + \langle t+1, j+1 \rangle = \sum_{j \subset t} \langle t+1, j \rangle + \sum_{j-1 \subset t} \langle t+1, j \rangle$. By Observation 5, this equals $\sum_{j \subset t-1} \langle t+1, j \rangle + 2 \sum_{j \subset t,\, j-1 \subset t} \langle t+1, j \rangle$. But ($j \subset t$ and $j-1 \subset t$) is equivalent to ($j \subset t$ and $2^{\lambda} \nmid j$). This proves Proposition 4. □

To prove Proposition 5, we apply Observation 1 to each block of A_t. Thus the i row (column) of A_t is replaced now by two rows (columns) which we denote by i, i^*. The i, j blocks of A_t (being $[t, i, j]$ iff $t \supset i$, $t \supset j$, $j \supset i$ and $b(j-i) = \{a_1, \cdots, a_{\nu}\}$ for some $\nu \geqq 0$) are replaced by

$[t+1, i+1, j+1]$	0
$[t+1, i, j+1]$	$[t+1, i, j]$

.

A zero block is replaced by

0	0
0	0

with the appropriate dimensions. The resulting matrix is called B_t; it is equal to A_t but described in a different way. B_t is, to sum up, a block matrix whose rows and columns are indexed by all i, i^* satisfying $i \subset t$. The only nonzero blocks in B_t are

$$\left.\begin{aligned} &B_t(i, j) = [t+1, i, j] \\ &B_t(i, j^*) = [t+1, i, j+1] \\ &B_t(i^*, j^*) = [t+1, i+1, j+1] \end{aligned}\right\} \text{ iff } \quad j \supset i,\ b(j-i) = \{a_1, \cdots, a_{\nu}\} \text{ for some } \nu \geqq 0.$$

We want to define nonsingular matrices P_t, Q_t such that in $C_t = P_t B_t Q_t$ the only nonzero blocks are, for $i, j \subset t$,

$$\left.\begin{aligned} &(i \neq 0) \quad C_t(i, j) = [t+1, i, j] \\ &(j \neq t) \quad C_t(i^*, j^*) = [t+1, i+1, j+1] \end{aligned}\right\} \text{ iff } \quad j \supset i,\ b(j-i) = \{a_1, \cdots, a_{\nu}\} \text{ for some } \nu \geqq 0,$$

$$C_t(0, j) = [t+1, 0, j] \quad \text{iff} \quad b(j) = \{a_1, \cdots, a_{\nu}\} \text{ for some } \nu \geqq 0 \text{ and } 2^{\lambda} | j,$$

$$C_t(i^*, t^*) = [t+1, i+1, t+1] \quad \text{iff} \quad b(i) = \{a_{\nu}, \cdots, a_{\tau}\} \text{ for some } \nu \geqq 1 \text{ and } 2^{\lambda} | (i+1),$$

$$C_t(0, t^*) = [t+1, 0, t+1].$$

The submatrix of C_t spanned by all $j \subset t$ with $2^\nu \| j$, $0 \leqq \nu < \lambda$, is equal to $A_{t+1-2^{\nu+1}}(2^{\nu+1}, 2^\nu, 2^\nu)$. To see this, we set a one-to-one correspondence between all $j' \subset t+1-2^{\nu+1}$ and all $j \subset t$ with $2^\nu \| j$, given by $j = j' + 2^\nu$. This shows the equality between these matrices. Also the submatrix generated by all j^* with $j \subset t$, $2^\nu \| (j+1)$, $0 \leqq \nu < \lambda$, equals $A_{t+1-2^{\nu+1}}(2^{\nu+1}, 2^\nu, 2^\nu)$. Here we correspond $j' \subset t+1-2^{\nu+1}$ to $j = j' + 2^\nu - 1$, $j \subset t$.

The remaining direct summand of C_t is the one indexed by all $j \subset t$ with $2^\lambda | j$, and by all j^* with $j \subset t$, $2^\lambda | (j+1)$. This submatrix is equal to A_{t+1}: Use the correspondence, to $i \subset t+1-2^\lambda$ assign $j = i \subset t$, and to $i \subset t+1$ with $2^\lambda \| i$ assign $j^* = (i-1)^*$ (note that $i - 1 \subset t$). This correspondence shows that this submatrix is really equal to A_{t+1}. Thus, if we can find nonsingular matrices P_t, Q_t so that $P_t B_t Q_t = C_t$, then Proposition 5 is established and therefore also the main theorem.

The matrices P_t, Q_t are defined inductively. Reminding the reader that $b(t) = \{a_1, \cdots, a_\tau\}$ with $a_1 > \cdots > a_\tau \geqq 0$, we do the induction on τ. For $\tau = 0$, i.e. $t = 0$,

$$A_0 = [0, 0, 0],$$

$$A_1 = B_0 = C_0 = \begin{array}{|c|c|} \hline [1,1,1] & 0 \\ \hline [1,0,1] & [1,0,0] \\ \hline \end{array}$$

and so P_0, Q_0 are defined to be identity matrices.

In the general case denote 2^{a_1} by δ, and $s = t - \delta$. We define L_t, K_t to be block matrices, indexed by all i, i^* where $i \subset t$. The only nonzero blocks in these matrices are the $(j+\delta, j^*)$ blocks $(j \subset s)$, which are $[t+1, j+\delta, j+1]_l$ and $[t+1, j+\delta, j+1]_k$ respectively.

Except for the cases $t = 2^\lambda - 1$, which will be dealt with later, we define

$$P_t = \begin{array}{|c|c|} \hline P_s(\delta,\delta) & \\ \hline & P_s(\delta, 0) \\ \hline \end{array} (I + L_t), \qquad Q_t = (I + K_t) \begin{array}{|c|c|} \hline Q_s(\delta,\delta) & \\ \hline & Q_s(\delta, 0) \\ \hline \end{array}.$$

Note that P_t depends on n, l, t only, and Q_t on n, k, t and so $P_s(x, y)(Q_s(x, y))$ results on replacing n by $n - x$ and l by $l - y$ (k by $k - y$), in $P_s(Q_s$ resp.).

To calculate the product $P_t B_t Q_t$ we start by working out

$$(I + L_t) B_t (I + K_t) = B_t + L_t B_t + B_t K_t + L_t B_t K_t.$$

The only nonzero blocks in $L_t B_t$ are $(i+\delta, j^*)$ blocks with $i \subset s$, $j \subset t$, $i \subset j$, $b(j-i) = \{a_1, \cdots, a_\nu\}$, $(0 \leqq \nu \leqq \tau)$. To find out what this block is we have to make the following product:

$$[t+1, i+\delta, i+1]_l [t+1, i+1, j+1] = [t+1, i+\delta, j+1] \cdot \binom{d+i+\delta-j-1}{\delta-1}.$$

The binomial coefficient is odd iff

$$\delta | (d+i-j).$$

We are assuming in Proposition 5 that $t < 2^\alpha$, where $2^\alpha \| d$, so $a_1 < \alpha$ and $\delta | d$. Hence, the condition is equivalent to $\delta | (j-i)$; but $j - i = 2^{a_1} + \cdots + 2^{a_h}$ and this is equivalent to $h = 0, 1$. Therefore, the only nonzero blocks in $L_t B_t$ are: for $j \subset s$ the $(j+\delta, j^*)$ block is $[t+1, j+\delta, j+1]$, and the $(j+\delta, (j+\delta)^*)$ block is $[t+1, j+\delta, j+\delta+1]$.

Similarly, the only nonzero blocks in B_tK_t are: for $j \subset s$, the (j, j^*) block is $[t+1, j, j+1]$ and the $(j+\delta, j^*)$ block is $[t+1, j+\delta, j+1]$. Therefore, in $L_tB_t + B_tK_t$ the only nonzero blocks are: for $j \subset t$, the (j, j^*) block is $[t+1, j, j+1]$.

It is easy to check that $L_tB_tK_t = 0$.

Note that the submatrix of B_t consisting of all $i+\delta$, $(i+\delta)^*$ rows and j, j^* columns with $i, j \subset s$ is equal to $B_s(\delta, 0, \delta)$, and so

$$(I+L_t)B_t(I+K_t) = \Lambda_t + \begin{array}{|c|c|}\hline 0 & 0 \\ \hline B_s(\delta, 0, \delta) & 0 \\ \hline \end{array},$$

where the only nonzero blocks in Λ_t are the (j, j) block $[t+1, j, j]$ and the (j^*, j^*) block $[t+1, j+1, j+1]$ for all $j \subset t$. Note also that $(I+L_t)\Lambda_t(I+K_t) = \Lambda_t$ (details are easy and are omitted) and so in the inductive process of defining P_t, Q_t we have $P_t\Lambda_tQ_t = \Lambda_t$.

By definition of Λ_t

$$\Lambda_t = \begin{array}{|c|c|}\hline \Lambda_s(\delta, \delta, \delta) & \\ \hline & \Lambda_s(\delta, 0, 0) \\ \hline \end{array},$$

and so

$$P_tB_tQ_t = \begin{array}{|c|c|}\hline P_s(\delta, \delta) & \\ \hline & P_s(\delta, 0) \\ \hline \end{array} \left\langle \Lambda_t + \begin{array}{|c|c|}\hline & \\ \hline B_s(\delta, 0, \delta) & \\ \hline \end{array} \right\rangle \begin{array}{|c|c|}\hline Q_s(\delta, \delta) & \\ \hline & Q_s(\delta, 0) \\ \hline \end{array}$$

$$= A_t + \begin{array}{|c|c|}\hline 0 & 0 \\ \hline C_s(\delta, 0, \delta) & 0 \\ \hline \end{array}.$$

In the last equality we made use of the fact that $P_s\Lambda_sQ_s = \Lambda_s$ and $P_sB_sQ_s = C_s$. It can be checked now that the only nonzero blocks $P_tB_tQ_t$ are given by: for $i, j \subset t$,

$$\left.\begin{array}{ll} (i \neq 0) & P_tB_tQ_t(i, j) = [t+1, i, j] \\ (j \neq t) & P_tB_tQ_t(i^*, j^*) = [t+1, i+1, j+1] \end{array}\right\} \text{ iff } j \supset i,\ b(j-i) = \{a_1, \cdots, a_\nu\} \text{ for some } \nu \geqq 0,$$

$$P_tB_tQ_t(0, j) = [t+1, 0, j] \quad \text{iff} \quad b(j) = \{a_1, \cdots, a_\nu\} \text{ with } \nu \geqq 0,\ 2^\mu \mid j,$$

$$P_tB_tQ_t(i^*, t^*) = [t+1, i+1, t+1] \quad \text{iff} \quad b(i) = \{a_\nu, \cdots, a_\tau\} \text{ with } \nu \geqq 1,\ 2^\mu \mid (i+1),$$

$$P_tB_tQ_t(0, t^*) = [t+1, 0, t+1],$$

where μ is defined by $2^\mu \| (s+1)$.

Since we assumed that t is different from $2^\lambda - 1$, it follows that $\mu = \lambda$, and so $P_tB_tQ_t = C_t$ as we wanted.

So assume $t = 2^\lambda - 1$ and so $\mu = \lambda - 1$ and $s = 2^\mu - 1$. In this case we define X_t (resp. Y_t) as we define P_t (resp. Q_t) in the general case. The only way $X_tB_tY_t$ differs from

C_t in this case is that it has the added nonzero $(0, j)$ blocks with $b(j)=\{a_1, \cdots, a_\nu\}$, $\nu \geqq 0$ and $2^\mu \| j$ and the (i^*, t^*) blocks with $b(i)=\{a_\nu, \cdots, a_\tau\}$ with $\nu \geqq 1$, $2^\mu \| (i+1)$. The only block of the first kind is the $(0, \delta)$ block which equals $[t+1, 0, \delta]$ and of the second kind, the (s^*, t^*) block, being $[t+1, s+1, t+1]$.

We define the matrices E_t(resp. F_t) as block matrices indexed by all i, i^* with $i \subset t$, and the only nonzero block being the $(s^*, 0)$ block which equals $[2^\lambda, 2^\mu, 0]_l$ (resp.$[2^\lambda, 2^\mu, 0]_k$). We define $P_t=(I+E_t)X_t$ and $Q_t=Y_t(I+F_t)$ and check that $P_tB_tQ_t=C_t$, as desired. This completes the proof of the main theorem.

A rank formula over $\mathbb{Z}_3$.

THEOREM 2. *The rank of $M_{n,l,l+1}$ over $\mathbb{Z}_3$ is*

$$\sum_{j \geqq 0} \binom{n-2j-1}{l-j}.$$

For $n \geqq 2l+1$ this equals

$$\sum_{j \geqq 0} \binom{n}{l-3j} - \sum_{j \geqq 0} \binom{n}{l-3j-2}.$$

Proof. Let F be a set of nonnegative integers; then we set $w(F)=\sum_{x \in F} 2^x$, (of course, $w=b^{-1}$). Let $X=\{1, \cdots, n\}$ be our base set. We show that $\mathscr{F}=\{F \in X^{(l)} | w(F) < 2^{n+2}/3\}$ is an independent set of rows. Since $\mathscr{F}=\{F \in X^{(l)} | n \notin F\} \cup \{F \in X^{(l)} | n \in F,\ (n-1) \notin F,\ (n-2) \notin F\} \cup \{F \in X^{(l)} | n \in F,\ (n-1) \notin F,\ (n-2) \in F, (n-3) \notin F, (n-4) \notin F\} \cup \cdots$, and this union is a disjoint union, $|\mathscr{F}|=\sum_{j \geqq 0} \binom{n-2j-1}{l-j}$ and this shows that the rank is at least this big. We prove that $\mathscr{F}$ is an independent set of rows by induction on n. For any n and $l=0, n-1$ this is clear. To perform the inductive step, define $Y=\{1, \cdots, n-2\}$,

$$\mathscr{B}_1=\{B \in Y^{(l-1)} | w(B) < 2^n/3\},$$

$$\mathscr{B}_2=\{B \in Y^{(l-1)} | w(B) > 2^n/3\}.$$

If $\mathscr{F}$ is dependent, this means that there is a function $f: \mathscr{F} \to \mathbb{Z}_3$, so that

$$\forall A \in X^{(l+1)} \quad \sum_{\substack{F \subset A \\ F \in \mathscr{F}}} f(F)=0$$

For $B \in \mathscr{B}_2$, let $A=B \cup \{n-1, n\}$, to obtain

$$f(B \cup \{n-1\})=0 \quad \forall B \in \mathscr{B}_2.$$

For $B \in \mathscr{B}_1$, $A=B \cup \{n-1, n\}$ we get $f(B \cup \{n-1\})+f(B \cup \{n\})=0$.
For $C \in Y^{(l)}$, let $A=C \cup \{n\}$; then we get

$$\forall C \in Y^{(l)} \quad f(C)+\sum_{\substack{B \subset C \\ B \in \mathscr{B}_1}} f(B \cup \{n\})=0$$

and for $A=C \cup \{n-1\}$ we have

$$f(C)+\sum_{\substack{B \subset C \\ B \in Y^{(l-1)}}} f(B \cup \{n-1\})=0,$$

$$f(C)+\sum_{\substack{B \subset C \\ B \in \mathscr{B}_1}} f(B \cup \{n-1\})+\sum_{\substack{B \subset C \\ B \in \mathscr{B}_2}} f(B \cup \{n-1\})=0.$$

All these equalities easily imply

$$\forall C \in Y^{(l)} \quad \sum_{\substack{B \subset C \\ B \in \mathcal{B}_1}} f(B \cup \{n\}) = 0.$$

But this shows that in $M_{n-2,l-1,l}$, where the basic set is Y, the rows of $\mathcal{B}_1 = \{B \in Y^{(l-1)} | w(B) < 2^n/3\}$ are linearly dependent, and this contradicts the induction hypothesis.

For the reverse inequality we first make:

Observation 6. Let P be a $p \times q$ matrix, Q a $q \times r$ matrix and R an $r \times s$ matrix. If $PQR = 0$, then

$$\rho(P) + \rho(Q) + \rho(R) \geqq q + r.$$

Now we prove

$$\rho(M_{n,l,l+1}) = \sum_{j \geqq 0} \binom{n-1-2j}{l-j}.$$

For $l \geqq 2$ we have that over $\mathbb{Q}$

$$M_{n,l-2,l-1} \cdot M_{n,l-1,l} \cdot M_{n,l,l+1} = 3M_{n,l-2,l+1}$$

so over $\mathbb{Z}_3$,

$$M_{n,l-2,l-1} \cdot M_{n,l-1,l} \cdot M_{n,l,l+1} = 0$$

and so over $\mathbb{Z}_3$,

$$\rho(M_{n,l-2,l-1}) + \rho(M_{n,l-1,l}) + \rho(M_{n,l,l+1}) \leqq \binom{n}{l-1} + \binom{n}{l} = \binom{n+1}{l}.$$

The l.h.s. is

$$\begin{aligned} &\geqq \sum_{j \geqq 0} \binom{n-1-2j}{l-2-j} + \binom{n-1-2j}{l-1-j} + \binom{n-1-2j}{l-j} \\ &= \sum_{j \geqq 0} \binom{n+1-2j}{l-j} - \binom{n-1-2j}{l-1-j} = \sum_{j \geqq 0} \binom{n+1-2j}{l-j} - \sum_{j \geqq 1} \binom{n+1-2j}{l-j} \\ &= \binom{n+1}{l}. \end{aligned}$$

It follows that all inequalities are in fact equalities, which completes the proof of the first assertion.

The proof that for $n \geqq 2l+1$

$$\sum_{j \geqq 0} \binom{n-2j-1}{l-j} = \sum_{j \geqq 0} \binom{n}{l-3j} - \sum_{j \geqq 0} \binom{n}{l-3j-2}$$

is straightforward, by induction on l. This formula was presented just because it resembles the rank formula of Theorem 1.

REFERENCES

[1] W. FOODY AND A. HEDAYAT, *On theory and applications of BIB designs with repeated blocks*, Ann. Statist., 5 (1977), pp. 932–945.

[2] R. GRAHAM, S.-Y. R. LI AND W.-C. LI, *On the structure of t-designs*, this Journal, 1 (1980), pp. 8–14.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 4, December 1981

0196-5212/81/0204-0001 $01.00/0

FLUCTUATION RESULTS FOR MARKOV-DEPENDENT TRIALS*

R. B. NAIN† AND KANWAR SEN†

Abstract. Takacs (SIAM Rev., 21 (1979), pp. 222–228) obtained explicit results for the distributions of the number of changes in luck in n Bernoulli trials with probability $p(0<p<1)$ for success. In this paper, the corresponding results have been obtained for Markov-dependent trials.

1. Introduction. With probability $p(0<p<1)$ for success in Bernoulli trials, Takacs (1979) obtained some explicit results of fluctuation, in terms of the first passage probabilities. Here, we have extended Takacs's technique to obtain the corresponding results for Markov-dependent trials.

Consider the Markov-dependent random variables $X_0, X_1, \cdots, X_n$, each assuming the values ± 1, with transition probabilities, $n \geqq 1$,

$$(1) \qquad \begin{array}{c} \\ P[X_{n-1}=1] \\ P[X_{n-1}=-1] \end{array} \begin{array}{c} \begin{array}{cc} P[X_n=1] & P[X_n=-1] \end{array} \\ \begin{bmatrix} p_1 & q_1 \\ q_2 & p_2 \end{bmatrix} \end{array}, \qquad p_1+q_1=p_1+q_2=1.$$

For $n=0$, let $P[X_0=1]=1-P[X_0=-1]=\rho_1$. Let $S_0=0$; $S_n=X_1+\cdots+X_n$, $n \geqq 1$ and $D_n^+=\max\{S_0, S_1, \cdots, S_n\}$. The sequence $\{S_n, n=0, 1, \cdots\}$ represents the positions of the walker in a Markov-dependent random walk.

This walk can be realized by repeated tossings of two coins, A and B, where A has probability p_1 for heads and B has probability q_2 for heads. Following a head, coin A is tossed, and following a tail, coin B is tossed. S_n is the cumulative number of heads minus the cumulative number of tails in n tosses.

Given $X_0=i$ $(=+1$ or $-1)$, we introduce the random variables

(i) $w_n^i(a)$, the number of subscripts $r=1, 2, \cdots$ for which either $S_{r-1}=a<S_r$ or $S_{r-1}>a=S_r$,

(ii) $\gamma_n^i(a)$, the number of subscripts $r=1, 2, \cdots$ for which $D_n^+=a$,

(iii) $v_n^i(a)$, the number of subscripts $r=1, 2, \cdots$ for which $S_r=a$, and

(iv) $\mu_n^i(a)$, the number of subscripts $r=1, 2, \cdots$ for which either $S_{r-1}<a<S_{r+1}$ or $S_{r-1}>a>S_{r+1}$.

2. Preliminary results. First we consider the probabilities of $S_n=r$ (r an integer). For $i=+1, -1$, $j=+1, -1$, let $u_r(n)$ and $f_r(n)$ be the 2×2 matrices

$$u_r(n)=(u_r^{i,j}(n)), \qquad f_r(n)=(f_r^{i,j}(n)),$$

where

$$u_r^{i,j}(n)=P\{S_n=r, X_n=j|X_0=i\},$$

$$f_r^{i,j}(n)=\begin{cases} P\{S_n=r, S_1<r, \cdots S_{n-1}<r, X_n=j|X_0=i\} & \text{if } r>0, \\ P\{S_n=r, S_1>r, \cdots S_{n-1}>r, X_n=j|X_0=i\} & \text{if } r<0. \end{cases}$$

Also, let

$$u_r^i(n)=u_r^{i,+1}(n)+u_r^{i,-1}(n).$$

Obviously, for $r>0$

$$f_r^{i,j}(n)\equiv f_r^{i,+1}(n)\equiv f_r^i(n), \qquad f_{-r}^{i,j}(n)\equiv f_{-r}^{i,-1}(n)\equiv f_{-r}^i(n).$$

* Received by the editors April 18, 1980, and in final form March 10, 1981.

† Department of Mathematical Statistics, University of Delhi, Delhi, India.

Define

$$U_r(r) = \sum_{n=0}^{\infty} u_r(n)t^n, \qquad F_r(t) = \sum_{n=1}^{\infty} f_r(n)t^n.$$

For a correlated random walk (1) Nain and Sen (1980) designed the matrices $\begin{bmatrix} p_1 & 0 \\ q_2 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & q_1 \\ 0 & p_2 \end{bmatrix}$, the matrices of the probabilities of "one-step" transition from a position to its right (i.e., from k to $k+1$) and left (i.e., from k to $k-1$), respectively. On multiplying these matrices by s and s^{-1} respectively, and then adding, we get

$$P(s) = \begin{bmatrix} p_1 s & q_1 s^{-1} \\ q_2 s & p_2 s^{-1} \end{bmatrix}. \tag{2}$$

From here it is straightforward to show that

$$U_r(t) = \text{coefficient of } s^r \text{ in } \sum_{n=0}^{\infty} [tP(s)]^n = [I - tP(s)]^{-1}$$

and, therefore

$$\begin{aligned} U_r(t) &= \frac{(Wp_1/p_2)^{r/2}}{p_1 t(1-W)} \begin{bmatrix} (Wp_1/p_2)^{1/2} - p_1 tW & q_1 tWp_1/p_2 \\ q_2 t & (Wp_1/p_2)^{1/2} - p_1 t \end{bmatrix}, \\ U_{-r}(t) &= \frac{(Wp_2/p_1)^{r/2}}{p_2 t(1-W)} \begin{bmatrix} (Wp_2/p_1)^{1/2} - p_2 t & q_1 t \\ q_2 tWp_2/p_1 & (Wp_2/p_1)^{1/2} - p_2 tW \end{bmatrix} \end{aligned} \tag{3}$$

for $r \geqq 1$, and

$$U_0(t) = \frac{(Wp_1/p_2)^{1/2}}{p_1 t(1-W)} \begin{bmatrix} 1 - t(Wp_1p_2)^{1/2} & q_1 t(Wp_1/p_2)^{1/2} \\ q_2 t(Wp_2/p_1)^{1/2} & 1 - t(Wp_1p_2)^{1/2} \end{bmatrix}, \tag{4}$$

where

$$W = \frac{1 - \{1 - 4p_1p_2t^2/(1+\delta t^2)^2\}^{1/2}}{1 + \{1 - 4p_1p_2t^2/(1+\delta t^2)^2\}^{1/2}}, \qquad \delta = p_1 - q_2.$$

The generating functions of the probabilities of first passage from origin to position r can be found by the following obvious relations, for $r > 0$

$$\begin{aligned} F_r^{+1}(t) &= \frac{U_r^{+1,j}(t)}{U_0^{+1,j}(t)} = [F_1^{+1}(t)]^r, \qquad F_r^{-1}(t) = \frac{U_r^{-1,j}(t)}{U_0^{+1,j}(t)} = F_1^{-1}(t)[F_1^{+1}(t)]^{r-1}, \\ F_{-r}^{+1}(t) &= \frac{U_{-r}^{+1,j}(t)}{U_0^{-1,j}(t)} = F_{-1}^{+1}(t)[F_{-1}^{-1}(t)]^{r-1}, \qquad F_{-r}^{-1}(t) = \frac{U_{-r}^{-1,j}(t)}{U_0^{-1,j}(t)} = [F_{-1}^{-1}(t)]^r \end{aligned} \tag{5}$$

where

$$\begin{aligned} p_2 F_1^{+1}(t) = p_1 F_{-1}^{-1}(t) &= \frac{1 + \delta t^2 - \{(1+\delta t^2)^2 - 4p_1p_2t^2\}^{1/2}}{2t}, \\ q_1 F_1^{-1}(t) = q_2 F_{-1}^{+1}(t) &= p_2 F_1^{+1}(t) - \delta t. \end{aligned} \tag{6}$$

The results at (5) and (6) have also been obtained by Jain (1971) by the method of difference equations.

Now, we introduce the random variable $\eta^i(k)$, $k \geqq 0$, such that

$$\eta^i(k) = \inf\{n, S_n = k \text{ and } n \geqq 0 | X_0 = i\}.$$

Clearly, $\eta^i(k)=\infty$, if $S_n \neq k$ for all $n \geqq 0$, and $P[\eta^i(k)=n]$ and $P[\eta^i(-k)=n]$ are the coefficients of t^n in $F_k^i(t)$ and $F_{-k}^i(t)$, respectively. Using the result of Jain (1971), we get

$$P[\eta^{+1}(k)=n]=\sum_{h=0}^{k}\binom{k}{h}p_1^{k-h}q_1^h\delta(k,h-k), \tag{7}$$

where

$$\delta(k,n)=\frac{(-\delta_2)^{n+k}}{(4p_2q_1)^k}\sum_{g=0}^{2k}\frac{\binom{2k}{g}(-1)^g\Gamma(k+1-g/2)}{\Gamma(n+k+1)\Gamma(1-n-g/2)}F(-g/2,-n-k;1-n-g/2;\delta_1/\delta_2)$$

and

$$\delta_1=\frac{\delta^2}{\delta_2}=(\sqrt{p_1p_2}-\sqrt{q_1q_2})^2.$$

It can now be obtained easily that

$$P[\eta^i(k)\leqq n]=P[S_n\geqq k|X_0=i]+\sum_m P[S_{n-m}<0|X_0=1]P[\eta^i(k)=m].$$

Writing generating functions

$$H_k^i(t)=\sum_{n=k}^{\infty}P[\eta^i(k)\leqq n]t^n, \quad \text{etc.},$$

we have

$$\begin{aligned}H_k^i(t)&=\sum_{r=k}^{\infty}U_r^i(t)+F_k^i(t)\sum_{r=1}^{\infty}U_{-r}^{+1}(t),\\ H_{-k}^i(t)&=\sum_{r=k}^{\infty}U_{-r}^i(t)+F_{-k}^i(t)\sum_{r=1}^{\infty}U_r^{-1}(t).\end{aligned} \tag{8}$$

Further, it is observed that the random variables $\eta^i(1)$, $\eta^{+1}(2)-\eta^i(1)$, $\eta^{+1}(3)-\eta^{+1}(2),\cdots,\eta^{+1}(k)-\eta^{+1}(k-1)$ are independent and identically distributed, except $\eta^i(1)$, whose distribution depends on i. Then

$$\begin{aligned}&\sum_{n_1+\cdots+n_k\leqq n}P[\eta^i(1)=n_1]P[\eta^{+1}(2)-\eta^i(1)=n_2]\cdots P[\eta^{+1}(k)-\eta^{+1}(k-1)=n_k]\\ &=\sum_{n_1+\cdots+n_k\leqq n}P[\eta^i(1)=n_1]P[\eta^{+1}(1)=n_2]\cdots P[\eta^{+1}(1)=n_k]\\ &=P[\eta^i(k)\leqq n]\end{aligned} \tag{9}$$

for $i\leqq k\leqq n$. This relation is similar to that of (10) in Takacs (1979) and forms the basis of the proofs of further results of this paper.

3. The distribution of $w_n^i(a)$. We shall prove the following:

THEOREM 1. *For $k\geqq 1$ and $a\geqq 0$, we have*

$$P[w_n^i(a)\geqq k]=\left(\frac{q_1}{q_2}\right)^{[k/2]}\left(\frac{p_2}{q_1}\right)^{k-1}\sum_{g=0}^{k-1}\binom{k-1}{g}\left(-\frac{\delta}{p_2}\right)^g P[\eta^i(a+k-g)\leqq n-g]. \tag{10}$$

Proof. Let us denote by $\theta_1,\theta_2,\cdots,\theta_k$, the successive values of $r(\geqq 0)$ for which S_r alternately takes the values $a+1$ and a. Then $S_r=a+1$ for $r=\theta_1,\theta_3,\cdots$ and $S_r=a$ for

$r=\theta_2, \theta_4, \cdots$. The random variables $\theta_1, \theta_2-\theta_1, \cdots, \theta_k-\theta_{k-1}, \cdots$ are independent and have distributions

$$
\begin{aligned}
&P[\theta_1=m|X_0=i]=P[\eta^{i}(a+1)=m],\\
&P[\theta_2-\theta_1=m|_1X_0=1]=P[\theta_4-\theta_3=m|_3X_0=1]\\
&\qquad=\cdots=P[\eta^{+1}(-1)=m]=\frac{q_1}{q_2}P[\eta^{-1}(1)=m],\\
&P[\theta_3-\theta_2=m|_2X_0=-1]=P[\theta_5-\theta_4=m|_4X_0=-1]=\cdots=P[\eta^{-1}(1)=m]
\end{aligned}
\tag{11}
$$

for $m\geqq 1$, where ${}_gX_0=i$ stands for that trial or the value of X (+1 or −1) at the end of which θ_g occurs.

$$
P[\eta^{-1}(1)\geqq 2]=\frac{p_2}{q_1}P[\eta^{+1}(1)\geqq 2],\qquad P[\eta^{-1}(1)=1]=\frac{p_2}{q_1}P[\eta^{+1}(1)=1]-\frac{\delta}{q_1}. \tag{12}
$$

Clearly, $w_n^i(a)\geqq k$ if and only if $\theta_1+(\theta_2-\theta_1)+\cdots+(\theta_k-\theta_{k-1})\leqq n$, given $X_0=i$. Therefore

$$
P[w_n^i(a)\geqq k]=P[\theta_k\leqq n|X_0=i]. \tag{13}
$$

By (9) and (12), we have

$$
\begin{aligned}
&P[\theta_1+(\theta_2-\theta_1)+\cdots+(\theta_k-\theta_{k-1})=m|X_0=i]\\
&\quad=\left(\frac{q_1}{q_2}\right)^{[k/2]}\sum_{m_1+\cdots+m_k=m}P[\eta^{i}(a+1)=m_1]P[\eta^{-1}(1)=m_2]\cdots P[\eta^{-1}(1)=m_k]\\
&\quad=\left(\frac{q_1}{q_2}\right)^{[k/2]}\left(\frac{p_2}{q_1}\right)^{k-1}\sum_{g=0}^{k-1}\binom{k-1}{g}\left(-\frac{\delta}{p_2}\right)^{g}P[\eta^{i}(a+k-g)=m-g].
\end{aligned}
\tag{14}
$$

Thus by (13) and (14) we obtain (10).

For $p_1=q_2=p(=1-q)$, the distribution of $w_n^i(a)$ reduces to the distribution of $w_n(a)$ for Bernoulli trials (Takacs (1979)).

4. The distribution of $\gamma_n^i(a)$.

THEOREM 2. *For $k\geqq 1$ and $a\geqq 1$, we have*

$$
P[\gamma_n^i(a)\geqq k]=p_2^{k-1}\sum_{g=0}^{k-1}\binom{k-1}{g}\left(-\frac{\delta}{p_2}\right)^{g}P[\eta^{i}(a+k-g-1)\leqq n-k+1-g]. \tag{15}
$$

Proof. Let $\theta_1', \theta_1'+\theta_2', \cdots, \theta_1'+\cdots+\theta_k', \cdots$ be the successive subscripts of $r=1, 2, \cdots$ for which $D_n^+=a$. Then

$$
P[\gamma_n^i(a)\geqq k]=P[\theta_1+\cdots+\theta_k\leqq n|X_0=i]. \tag{16}
$$

The random variables $\theta_1', \theta_2', \cdots, \theta_k', \cdots$ are independent and each has distribution

$$
P[\theta_g'=m|_{g-1}X_0=+1]=q_1P[\eta^{-1}(1)=m-1], \tag{17}
$$

except θ_1', which has distribution

$$
P[\theta_1'=m|X_0=i]=P[\eta^{i}(a)=m]. \tag{18}
$$

Then

$$P[\theta_1' + \cdots + \theta_k' = m | X_0 = i] = p_2^{k-1} \sum_{g=0}^{k-1} \binom{k-1}{g} \left(-\frac{\delta}{p_2}\right)^g P[\eta^i(a+k-g-1) = m-k+1-g]. \tag{19}$$

Now, by (16) and (19) we get (15).

It seems that, for $p_1 \neq p_2$, the distribution of $v_n^i(a)$ and $\mu_n^i(a)$ are not easily obtainable by the technique being followed here. However, these distributions have been found by the same technique for symmetrical Markov-dependent trials, i.e., $p_1 = p_2 = p$. Hereafter we shall use the above notation, but for $p_1 = p_2 = p$ only.

5. The distribution of $v_n^i(a)$. Here we shall prove the following:

THEOREM 3. *For $k \geqq 1$, we have*

$$P[v_n^i(a) \geqq k] = \begin{cases} \left(\frac{p}{k}\right)^k \sum_{g=0}^{k} \binom{k}{g} \left(-\frac{c}{p}\right)^g P[\eta^{+1}(k-g) \leqq n-k-g] & \text{if } a = 0, \\ \left(\frac{p}{q}\right)^{k-1} \sum_{g=0}^{k-1} \binom{k-1}{g} \left(-\frac{c}{p}\right)^g P[\eta^i(a+k-1-g) \leqq n-k+1-g] & \text{if } a \geqq 1, \end{cases} \tag{20}$$

where $c = p - q$.

Proof. Let $\alpha_1, \alpha_1+\alpha_2, \cdots, \alpha_1+\cdots\alpha_k, \cdots$ be the successive values of $r = 1, 2, \cdots$ for which $S_r = a$. Then

$$P[v_n^i(a) \geqq k] = P[\alpha_1 + \cdots + \alpha_k \leqq n | X_0 = i] \tag{21}$$

for $k \geqq 1$ and $n \geqq 1$.

For $a = 0$, the variables $\alpha_1, \alpha_2, \cdots, \alpha_k, \cdots$ are independent and each has distribution

$$P[\alpha_g = m |_{g-1} X_0 = i] = P[\eta^{-1}(1) = m]. \tag{22}$$

Accordingly, by (9), (12) and (22), we get

$$P[\alpha_1 + \cdots + \alpha_k = m | X_0 = i] = \left(\frac{p}{q}\right)^k \sum_{g=0}^{k} \binom{k}{g} \left(-\frac{c}{p}\right)^g P[\eta^{+1}(k-g) = m-k-g]. \tag{23}$$

Now by (21) and (23) we get (20) for $a = 0$.

For $a > 0$, the variables $\alpha_1, \alpha_2, \cdots, \alpha_k, \cdots$ are again independent and each has distribution (22), except α_1. The distribution of α_1 in this case is

$$P[\alpha_1 = m | X_0 = i] = P[\eta^i(a) = m]. \tag{24}$$

Then

$$P[\alpha_1 + \cdots + \alpha_k = m | X_0 = i] = \left(\frac{p}{q}\right)^{k-1} \sum_{g=0}^{k-1} \binom{k-1}{g} \left(-\frac{c}{p}\right)^g P[\eta^i(a+k-1-g) = m-k+1-g]. \tag{25}$$

By (21) and (25), we prove (20) for $a > 0$.

6. The distribution of $\mu_n^i(a)$.

THEOREM 4. *For $k \geqq 0$, we have*

$$P[\mu_n^i(a) \geqq k] = \begin{cases} \left(\frac{p}{q}\right)^k \sum_{g=0}^{k} \binom{k}{p}\left(-\frac{c}{p}\right)^g P[\eta^{+1}(2k-2g) \leqq n-g] & \text{if } a=0, \\ \left(\frac{p}{q}\right)^{k-1} \sum_{g=0}^{k-1} \binom{k-1}{g}\left(-\frac{c}{p}\right)^g P[\eta^i(a+2k-2g-1) \leqq n-g+1] & \text{if } a>0. \end{cases} \tag{26}$$

Proof. Denote by $\alpha_1^*, \alpha_1^*+\alpha_2^*, \cdots, \alpha_1^*+\cdots+\alpha_k^*, \cdots$ the successive values of $r=1, 2, \cdots$ for which either $S_{r-1} < a < S_{r+1}$ or $S_{r-1} > a > S_{r+1}$. Then, we have

$$P[\mu_n^i(a) \geqq k] = P[\alpha_1^* + \cdots + \alpha_k^* \leqq n | X_0 = i] \tag{27}$$

for $n \geqq 1$ and $k > 1$.

First, for $a = 0$, the random variables $\alpha_1^*, \alpha_2^*, \cdots, \alpha_k^*, \cdots$ are independent and each has distribution

$$P[\alpha_g^* = m|_{g-1} X_0 = i] = P[\eta^{-1}(2) = m]. \tag{28}$$

Then

$$P[\alpha_1^* + \cdots + \alpha_k^* = m | X_0 = i] = \left(\frac{p}{q}\right)^k \sum_{g=0}^{k} \binom{k}{g}\left(-\frac{c}{p}\right)^g P[\eta^{+1}(2k-2g) = m-g]. \tag{29}$$

Now by (27) and (29) we get (26) for $a = 0$.

For $a > 0$, the variables $\alpha_1^*, \alpha_2^*, \cdots, \alpha_k^*, \cdots$ are still independent and each has distribution (28), except α_1^* which has distribution

$$P[\alpha_1^* = m | X_0 = i] = P[\eta^i(a+1) = m+1].$$

Thus

$$\begin{aligned} &P[\alpha_1^* + \cdots + \alpha_k^* = m | X_0 = i] \\ &\quad = \left(\frac{p}{q}\right)^{k-1} \sum_{g=0}^{k-1} \binom{k-1}{g}\left(-\frac{c}{p}\right)^g P[\eta^i(a+2k-2g-1) = m+1-g]. \end{aligned} \tag{30}$$

Hence by (27) and (30) we prove (26) for $a > 0$.

For $p = q = \frac{1}{2}$, the distributions of $v_n^i(a)$ and $\mu_n^i(a)$ reduce to the distributions of $v_n(a)$ and $\mu_n(a)$, for Bernoulli trials with $\frac{1}{2}$ as the probability of success (Takacs (1979)).

REFERENCES

J. N. DARROCH AND J. WHITFORD (1972), *Exact fluctuation results for Markov-dependent coin tossing*, J. Appl. Probab., 9, pp. 158–168.

G. C. JAIN (1971), *Some results in a correlated random walk*, Canad. Math. Bull., 14, pp. 341–347.

R. B. NAIN AND K. SEN (1980), *Transition probability matrices for correlated random walks*, J. Appl. Probab., 17, pp. 253–258.

L. TAKACS (1979), *Fluctuation problems for Bernoulli trials*, SIAM Rev., 21, pp. 222–228.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 4, December 1981

0196-5212/81/0204-0002 $01.00/0

WEIGHT ENUMERATORS OF NORMALIZED CODES*

STEPHEN M. GAGOLA, JR.†

Abstract. A linear code C over a finite field is normalized if it contains the all ones vector. If a normalized code C is also self-dual then its complete weight enumerator is invariant under the action of a linear group G, which is explicitly determined. The character of this representation is then used to calculate the Molien series for G.

Further restrictions on C may lead to larger finite linear groups containing G. It is determined here that if the field is not $GF(2)$ or $GF(4)$ then there are only finitely many linear groups containing G with the property that the only scalar matrices appearing are those already contained in G. In fact, if the characteristic is odd and $\tilde{G}$ is the unimodular subgroup of G, then the finite unimodular subgroups containing $\tilde{G}$ are contained in a unique, such maximal linear group. The classification of the finite simple groups is used for the proof of this last result.

1. Introduction. A linear code is defined to be normalized if it contains the all ones vector. In this paper the complete weight enumerator for a normalized self-dual code is shown to be invariant under the action of a linear group G, which is explicitly determined up to an isomorphism. The Molien series for G is also determined. Because there seems to be a fundamental difference between odd characteristic and characteristic two, the discussion of the construction of G is divided into two sections (§§ 4 and 5), and likewise for the construction of the Molien series (§§ 6 and 7).

A. M. Gleason [6], has determined a linear group G_0 under which the complete weight enumerator of a self-dual code is invariant. As reported in [15], the Molien series for G_0 has also been determined (unpublished) by A. M. Gleason, R. J. McEliece, E. R. Rodemich and H. C. Rumsey, Jr. Since G_0 appears as a subgroup of G, these results are recovered here. (In characteristic two $G_0 = G$, while in odd characteristic p, G is an extension of G_0 by an appropriate extraspecial p-group of exponent p.)

For small values of q, the complete ring of invariants of G_0 has been determined. The corresponding problem for G will not be considered here, however. For $q = 3$, the Molien series of G and generators for the corresponding ring of invariants are given in [19]. In that paper G is referred to as G_7, the linear group of degree 3 and order 2592. This group is also discussed in [16], where it is denoted by $\mathcal{G}_{2592}$.

In this paper, the Molien series for G and G_0 are given in unsimplified form, but for small values of q, $q \leqq 9$ are worked out explicitly in the appendix. The calculations there were originally done by hand in the case of $\Phi_{G_0}(X)$ for $q \leqq 9$ and $\Phi_G(X)$ for $q = 3$. These were later checked independently by using the HP9830A calculator, which also covered the remaining cases not done by hand. I wish to acknowledge here the help of my wife Gloria, who did most of the calculations.

Linear groups containing G (for odd characteristic) are discussed in § 8.

Much of what is done here carries over to the "Hermitian case" (when F has an automorphism of order two) and this will be the topic of a sequel.

2. Preparations from coding theory. Let F be a finite field. By a *code* in F^n we shall always mean a linear code, that is, a subspace of F^n. If C is a code in F^n, then $C^\perp$ denotes the dual code of C, i.e.,

$$C^\perp = \{v \in F^n \mid \sum v_i c_i = 0 \text{ for all } c \in C\}.$$

* Received by the editors April 24, 1980, and in final form February 2, 1981.
† Department of Mathematics, Texas A & M University, College Station, Texas 77843.

A code is *self-dual* if $C^{\perp} = C$ and is *normalized* if it contains the all ones vector $\mathbb{1}$. Clearly, if F^n contains a normalized self-dual code then n must be even and a multiple of the characteristic of F. If char $F = 2$, it is also readily checked that any self-dual code is necessarily normalized.

For each element a of F let X_a be an indeterminate. The complete weight enumerator for a code $C \subseteq F^n$ is defined by the equation

$$W_C = \sum X_{a_1} X_{a_2} \cdots X_{a_n},$$

where the sum extends over all vectors $(a_1, a_2, \cdots, a_n)$ in C. If $|F| = q$ is odd and C is self-dual, W_C is invariant under the action of a group G_0 of $q \times q$ matrices, which, as an abstract group, is isomorphic to $SL(2, F) \times J$ [6]. Here J is cyclic of order 2 or 4, according to whether $q \equiv 1$ or 3 mod 4, respectively. This group is contained in a larger matrix group G (constructed in §§ 4 and 5) which leaves invariant the complete weight enumerator of any normalized self-dual code. Matrix generators for these groups appear at the end of this section.

Starting with a normalized self-dual code C, it is possible to impose conditions on the complete weight enumerator, which imply that it is invariant under the action of a still larger matrix group. One obvious example of this is to require that the dimension n be a multiple of some fixed integer k. The corresponding extension of G is obtained by adjoining the scalar matrix εI, where ε is a primitive kth root of 1. Another example is to assume that the subspace C is setwise invariant under the action of the Galois group of F over some subfield. Here, the larger matrix group is obtained by forming the semidirect product of G with the Galois group, and so is not a simple "scalar extension."

Scalar extensions of G may be regarded as trivial. With this in mind, it becomes interesting to classify all finite subgroups H of $GL(q, \mathbb{C})$ which contain G and which contain no scalar matrices other than those already contained in G. For a given value of q different from 2 or 4, the matrix group G is primitive, and, as a consequence of a theorem of Jordan [13], there are only finitely many possibilities for H. For a modern treatment of Jordan's theorem see [11, Thm. 14.12], or [4, Thm. 30.3], where a slightly better bound for $|H: \mathbb{Z}(H)|$ is obtained.

For odd q the group G has the form $C \times \tilde{G}$, where C is cyclic of order 2 or 4 and is represented by scalar matrices. The group $\tilde{G}$ turns out to be perfect (equal to its commutator subgroup $\tilde{G}'$) when $q > 3$, and hence is the unimodular subgroup of G. When $q = 3$, $\tilde{G}$ is not perfect and $\tilde{G}'$ is the unimodular subgroup of G. The linear group $\tilde{G}'$ is still primitive and by the remarks above, is the unimodular subgroup of G for all odd q. For convenience we shall restrict our attention to the finite unimodular subgroups H of $GL(q, \mathbb{C})$ (and hence of $SL(q, \mathbb{C})$) containing $\tilde{G}'$. The result obtained in § 8 is that there is a unique largest such group H.

When $q = 2$ or 4, it turns out that G is similar to a group of monomial matrices, and because of this there are infinitely many possibilities for H. Indeed, when $q = 2$, G is a dihedral group of order 16. This matrix group is therefore contained in an infinite collection of larger matrix groups, each corresponding to an irreducible representation of a larger dihedral group.

The remarks following the next theorem will attempt to justify considering only *finite* subgroups of $GL(q, \mathbb{C})$ containing G. The theorem itself is an example of the type of result obtained using weight enumerators. N. J. A. Sloane attributes it to some unpublished work of Gleason [20].

THEOREM 2.1. *Let $k > 1$ be an integer, F a finite field with q elements and $C \subseteq F^{2n}$ a self-dual code. If the weights of all vectors in C are divisible by k, then one of the*

following holds:

(1) $(k, q) = (2, 2), (2, 4), (3, 3)$ *or* $(4, 2)$.

(2) $k = 2$, $q \geqq 8$ *is a power of* 2 *and* C *is equivalent to a direct sum of* n *copies of the code* $\{(a, a) \mid a \in F\}$.

(3) $k = 2$, $q \equiv 1 \bmod 4$ *and* C *is equivalent to a direct sum of* n *copies of the code* $\{(a, ca) \mid a \in F\}$, *where* $c \in F$ *and* $c^2 = -1$.

The proof of this result uses the Hamming weight enumerator, which is an invariant of the matrix group $\left\langle 1/\sqrt{q}\begin{pmatrix} 1 & q-1 \\ 1 & -1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix} \right\rangle$, where ε is a primitive kth root of 1. This matrix group may be viewed as an irreducible complex representation of degree 2 of an abstract group $G(k, q)$. When the parameters (k, q) are equal to one of the four possibilities given in (1), the group $G(k, q)$ is finite. In fact, we have the isomorphisms

$$G(2, 2) \simeq D_{16}, \qquad G(2, 4) \simeq D_{12},$$

$$G(3, 3) \simeq SL(2, 3) \Upsilon C_4, \qquad G(4, 2) \simeq GL(2, 3) \Upsilon C_8.$$

Here, C_n and D_n denote a cyclic and dihedral group of order n respectively, and Υ denotes a central product. The Molien series for the given two-dimensional representation of $G(k, q)$ is known in each case, and in fact generators for the full ring of invariants are known and listed in [2].

If $k = 2$ and either $q = 3$ or $q > 4$, the group $G(k, q)$ is infinite dihedral, and the ring of polynomial invariants relative to the given two-dimensional representation is generated by the single invariant $x^2 + (q-1)y^2$. For the remaining choices for (k, q) the group $G(k, q)$ is infinite, not dihedral, and the ring of polynomial invariants consists only of constants.

Thus, when $G(k, q)$ is infinite, the code is either highly restricted, or does not exist at all. Returning to the situation discussed before the statement of the theorem, it seems reasonable to consider only those conditions on the complete weight enumerator of a normalized self-dual code which lead to a *finite* subgroup of $GL(q, \mathbb{C})$, which contains G.

In order to describe matrix generators for G_0 and G, it is convenient to introduce some notation which describes the characters of F. Recall that a character of F is any homomorphism from the additive group of F into the multiplicative group $\mathbb{C}^\times$. Regarding $(F, +)$ as an abstract group, the characters of F coincide with the irreducible characters of the group $(F, +)$. The trivial map $F \to \{1\}$ is then the principal character of F. Let $GF(p)$ denote the prime subfield of F and let $\mu_0 : GF(p) \to \mathbb{C}^\times$ be the "standard character" given by $\mu_0(j) = \exp(2\pi ij/p)$. Let $\mathrm{tr} : F \to GF(p)$ be the trace map and define $\lambda : F \to \mathbb{C}^\times$ by $\lambda(a) = \mu_0(\mathrm{tr}(a))$. Thus, λ is a nonprincipal irreducible character of F. If $b \in F$, define $\lambda_b(a) = \lambda(ba)$ for all $a \in F$. Thus, λ_0 is the principal character of F, $\lambda_1 = \lambda$ and $\{\lambda_b \mid b \in F\}$ is the full set of irreducible characters of F. This notation will be retained throughout the entire paper.

In order to state the MacWilliams identity it is convenient to introduce some extra notation. If $f(X_1, \cdots, X_q)$ is a polynomial and $M = (m_{ij})$ is a $q \times q$ matrix, let $f \cdot M$ denote the polynomial $f(Y_1, \cdots, Y_q)$ where $Y_i = \sum m_{ij} X_j$. Finally, a square matrix whose rows and columns are naturally indexed by some set S will be called an $S \times S$ matrix.

THEOREM 2.2 (MacWilliams). *Let* $C \subseteq F^n$ *be any (linear) code and for each* $a \in F$ *let* M'_a *be the* $F \times F$ *matrix over* $\mathbb{C}$ *whose* (r, s) *entry is* $\lambda_a(rs)$. *Then, for every* $a \neq 0$ *in* F,

the weight enumerators W_C and $W_{C^\perp}$ are related by

$$W_{C^\perp} = \frac{1}{|C|} W_C \cdot M'_a.$$

In particular, if C is self-dual, then W_C is invariant under the action of $q^{-1/2}M'_a$.

The original version of this theorem appears in [14]. The proof of the version given above closely parallels the original, and will be omitted.

For each $a \in F$, let M_a, N_a, D_a and E_a be the $F \times F$ matrices with entries in $\mathbb{C}$ defined as follows:

$$(M_a)_{r,s} = q^{-1/2}\lambda_a(rs) \qquad (\text{thus } M_a = q^{-1/2}M'_a),$$
$$(N_a)_{r,s} = \delta_{ra,s},$$
$$(D_a)_{r,s} = \delta_{r,s}\lambda_a(r^2),$$
$$(E_a)_{r,s} = \delta_{r,s}\lambda_a(r).$$

Here $\delta_{r,s}$ denotes the Kronecker delta. By Theorem 2.2 we have $W_C = W_C \cdot M_a$ for every self-dual code C and any $a \in F$, $a \neq 0$. When $a = 0$, Theorem 2.2 does not apply, even though M_0 and N_0 are still defined. However, these matrices are singular, and will not be used.

Notice that the matrix M_a is ambiguous, as the sign of the square root $q^{-1/2}$ was not specified. We assume that some choice is made for this sign, and that this same choice is made simultaneously for all of the matrices M_a, $a \in F$.

Since $\sum c_i^2 = 0$ holds for every vector $(c_1, \cdots, c_n)$ of a self-dual code C, it readily follows that W_C is invariant under D_a for all $a \in F$. If in addition C is normalized then $\sum c_i = 0$ also holds, and W_C is invariant under E_a. That W_C is also invariant under N_a for $a \neq 0$, is a consequence of the next result (although this also follows easily from the definition of W_C and the linearity of C).

LEMMA 2.3. *For every $a, b, c, d \in F$ with $a \neq 0$ and $b \neq 0$ we have*

$$M_aM_b = N_{-ab^{-1}}, \qquad M_a^{-1}N_bM_a = N_{b^{-1}},$$
$$N_aN_b = N_{ab}, \qquad N_a^{-1}D_cN_a = D_{a^{-2}c},$$
$$D_cD_d = D_{c+d}, \qquad N_a^{-1}E_cN_a = E_{a^{-1}c},$$
$$E_cE_d = E_{c+d}.$$

These relations are easily checked and the proof will be omitted. The matrices M_a, N_a and D_a (for appropriate a) are discussed in [15], where they are denoted by T_1, T_4 and T_5 respectively.

From the remarks preceding the theorem, the weight enumerator of a self-dual code is an invariant of the matrix group $G_0 = \langle M_a, D_c \mid a, c \in F, a \neq 0\rangle$, while that of a normalized self-dual code is an invariant of $G = \langle M_a, D_c, E_d \mid a, c, d \in F, a \neq 0\rangle$. These groups are determined abstractly in §§ 4 and 5, and the Molien series for the given matrix representations of them are worked out in §§ 6 and 7.

3. Preparations from group theory. Most of the notation and terminology used is standard. The general references for group theory chosen here are Gorenstein's book [7] and Huppert's book [10], although there are many other good books. I. M. Isaacs' book [11] and L. Dornhoff's book [4] are used as references for representation theory. In particular, Dornhoff's book contains a complete character table for the group $SL(2, q)$ which will be referred to in § 5.

If G is a group and H is a subset of G, we will write $H \leqq G$ to mean that H is a subgroup of G. If g_α for $\alpha \in \Lambda$, are elements of G, let $\langle g_\alpha \mid \alpha \in \Lambda \rangle$ denote the subgroup generated by $\{g_\alpha \mid \alpha \in \Lambda\}$. If $H \leqq G$, a (right) transversal for H in G is a set T of coset representatives for H in G. Thus $G = \cup_{t \in T} Ht$, where the union is not redundant. The notation $|S|$ will always mean the cardinality of the set S. Hence, if T is a transversal for H in G then $|T| = |G:H|$, where $|G:H|$ denotes the index of H in G.

If $H, K \leqq G$ and G is the internal direct product of H and K, then $G = H \times K$. The core in G of a subgroup H, denoted by $\text{core}_G(H)$, is the largest normal subgroup of G that is contained in H. Hence,

$$\text{core}_G(H) = \bigcap_{x \in G} x^{-1}Hx.$$

An action of a group G on a set Ω is a function $\Omega \times G \to \Omega$, where $\alpha^1 = \alpha$ and $(\alpha^g)^h = \alpha^{gh}$. Here, the image of (α, g) is denoted by α^g. If H and N are groups, an action by automorphisms of H on N is an action which satisfies $(xy)^h = x^h y^h$ for $x, y \in N$ and $h \in H$.

If H acts on N by automorphisms, then the semidirect product $G = H \ltimes N$ may be formed as follows. As a set, G is $H \times N$, and the group multiplication is given by: $(h, n)(h', n') = (hh', n^{h'}n')$. It is easily checked that G is a group, and H and N will be regarded as subgroups of G via the natural embeddings. In particular, $n^h = h^{-1}nh$ so that the action of H on N becomes conjugation within G. The group H acts faithfully on N if the kernel of the action is trivial, that is $\mathbb{C}_H(N) = \{1\}$.

A representation ρ of degree n of a group is a homomorphism $\rho: G \to GL(n, \mathbb{C})$. The corresponding character χ afforded by this representation is the function $\chi: G \to \mathbb{C}$, defined by $\chi(g) = \text{trace}(\rho(g))$. Two representations ρ and ρ' of degree n are equivalent if there exists $t \in GL(n, \mathbb{C})$ such that $\rho(g) = t^{-1}\rho'(g)t$ for all $g \in G$. It is well known in representation theory that two representations are equivalent if and only if they afford the same character [11, Corollary 2.9]. A representation of degree one will always be identified with its character and will be called a linear character of G. If $G = H \times K$, and if λ and μ are linear characters of H and K respectively then $\lambda \# \mu$ denotes the linear character of G defined by $(\lambda \# \mu)(hk) = \lambda(h)\mu(k)$. If ρ is any representation of G then $\det \rho$ denotes the linear character whose value at $g \in G$ is $\det(\rho(g))$. It is convenient to regard any representation, or character, of a factor group G/N as a representation or character of G (with N in its kernel).

If ρ is a representation of $H \leqq G$ then ρ^G denotes the corresponding induced representation. If T is a transversal for H in G then $\rho^G(g)$ may be defined as a $T \times T$ matrix in blocked form, as follows. The (t, u) entry of $\rho^G(g)$ is $\rho^0(tgu^{-1})$, where $\rho^0(x) = \rho(x)$ for $x \in H$ and $\rho^0(x)$ is the zero matrix (of the appropriate size) otherwise. If ρ affords the character χ, then χ^G will denote the character afforded by ρ^G. The representation ρ^G does depend on the choice of T, but its character χ^G does not. In particular, ρ^G is well defined up to equivalence of representations.

The kernel of ρ^G is the largest normal subgroup of G which is contained in $\ker(\rho)$. Thus, $\ker \rho^G = \text{core}_G(\ker \rho) = \cap_{x \in G} x^{-1}(\ker \rho)x$.

If K is another subgroup of G, then by the Mackey decomposition, the restriction of ρ^G to K (denoted $\rho^G|_K$) is equivalent to a direct sum of representations of the form $(\rho^x|_{x^{-1}Hx \cap K})^K$, where x ranges over a set of double coset representatives for (H, K) in G. Here, ρ^x is the representation of $x^{-1}Hx$ defined by $\rho^x(x^{-1}hx) = \rho(h)$. If χ is the character afforded by ρ, let χ^x denote the character afforded by ρ^x. Hence, $\chi^x(x^{-1}hx) = \chi(h)$. A special case is worth pointing out. If $H \lhd G$, then χ^x is a character of H for all $x \in G$, and G acts transitively on the set $\{\chi^x \mid x \in G\}$. The inertia group of χ in G, denoted

$I_G(\chi)$, is the stabilizer of χ in G under this action. Thus, $I_G(\chi) = \{g \in G \mid \chi^g = \chi\}$. Since χ is a class function on H it follows that $H \leqq I_G(\chi)$. If χ is an irreducible character of H and $I_G(\chi) = H$, then χ^G is irreducible (see [11, Thm. 6.11]).

If χ is an irreducible representation of the normal subgroup N of G, say that χ is extendible to G if there exists a character $\hat{\chi}$ of G such that $\hat{\chi}|_N = \chi$. Clearly, if χ is extendible to G then $I_G(\chi) = G$. The converse is not true. However, it is convenient to record here the following extendibility theorem.

THEOREM 3.1. *Let χ be an irreducible character of the normal subgroup N of G. Assume $I_G(\chi) = G$. Then χ is extendible to G if either*

(1) *G/N is cyclic, or*

(2) *χ is extendible to all subgroups H of G, containing N, such that H/N is a Sylow p-subgroup of G/N for all prime divisors of $|G:N|$.*

(The first part of this theorem is [11, Corollary 11.22] and the second part is [11, Corollary 11.31].)

We remark, that in the second case of the theorem, if H/N is a Sylow p-subgroup of G/N and χ is extendible to H, then χ is extendible to $g^{-1}Hg$ for all $g \in G$. Hence, extendibility need only be checked for a single Sylow p-subgroup of G/N for each prime divisor p of $|G:N|$.

If χ is extendible to a character $\hat{\chi}$ of G, then $\hat{\chi}$ is not necessarily unique. However, every extension of χ to G has the form $\lambda\hat{\chi}$ where λ is a linear character of G with kernel containing N. If G/N is perfect, i.e., equal to its commutator subgroup, then the only choice for λ is the principal character 1_G, and $\hat{\chi}$ is unique.

The following theorem provides a uniqueness result for extensions of representations.

THEOREM 3.2. *Let χ be an irreducible character of the subgroup N of G afforded by ρ. If χ is extendible to a character $\hat{\chi}$ of G, then ρ is extendible to a unique representation of G affording $\hat{\chi}$.*

Proof. Let $\hat{\rho}$ be a representation affording $\hat{\chi}$. Then $\hat{\rho}|_N$ affords $\chi_N = \chi$, so $\hat{\rho}|_N$ is equivalent to ρ. Replacing $\hat{\rho}$ by an equivalent representation, we may assume $\hat{\rho}_N = \rho$. Suppose now $\tilde{\rho}$ is another representation of G affording $\hat{\chi}$, satisfying $\tilde{\rho}|_N = \rho$. Since $\tilde{\rho}$ and $\hat{\rho}$ afford the same character, there exists a nonsingular matrix t such that

$$\tilde{\rho}(g) = t^{-1}\hat{\rho}(g)t \quad \text{for all } g \in G.$$

Now $\tilde{\rho}|_N = \hat{\rho}|_N$ so t commutes with $\rho(x)$ for all $x \in N$, and this implies, by Schur's lemma, that t is a scalar matrix. Hence $\tilde{\rho} = \hat{\rho}$ and the theorem follows. □

4. Construction of G (odd characteristic). In this section, $F = FG(q)$, where q is a power of the prime p and $p \geqq 3$. Let V denote the two-dimensional row space over F, that is, $V = \{(a, b) \mid a, b \in F\}$ and define the group E as follows: As a set, E is $V \times GF(p)$. If (u, m) and (v, n) are in E, define

$$(u, m)(v, n) = (u + v, m + n + \operatorname{tr}(u_1v_2 - u_2v_1)).$$

Here $\operatorname{tr}: F \to GF(p)$ is the usual trace map, $u = (u_1, u_2)$ and $v = (v_1, v_2)$. It is readily checked that E is indeed a group and in fact is an extraspecial p-group of exponent p and of order q^2p. It will sometimes be convenient to identify the set $F \times F \times GF(p)$ with E.

Define an action of the special linear group $SL(2, F)$ on E as follows: For $g \in SL(2, F)$ and $e = (v, n) \in E$, define $e^g = (vg, n)$. Again, it is readily checked that this is an action, and since $\det g = 1$ for $g \in SL(2, F)$, the action is by automorphisms. Notice that $SL(2, F)$ acts on $E/\mathbb{Z}(E)$ and that this action may be identified with the natural

action of $SL(2, F)$ on V. Let $\tilde{G}$ denote the semidirect product $SL(2, F) \ltimes E$. Hence, as a set $\tilde{G}$ is $SL(2, F) \times E$, with the group multiplication defined by $(g, e)(h, f) = (gh, e^h f)$. As usual, $SL(2, F)$ and E will be regarded as subgroups of $\tilde{G}$, via the natural embeddings.

It is convenient to fix a notation for certain subgroups of $\tilde{G}$. Let $A_0 = \{0\} \times F \times \{0\} \leqq E$, and $\mathbb{Z} = \{0\} \times \{0\} \times GF(p) \leqq E$. Notice that A_0 is elementary abelian and $\mathbb{Z} = \mathbb{Z}(E)$. Let $A = A_0 Z = A_0 \times \mathbb{Z}$, so that A is also elementary abelian. Let $S = SL(2, F)$, and define

$$P = \left\{ \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \middle| a \in F \right\} \leqq S \quad \text{and} \quad H = \left\{ \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} \middle| a \in F^\times \right\} \leqq S.$$

Then H normalizes P and the group PH normalizes A_0, so PHA_0 is a group. The group $\mathbb{Z}$ is centralized by PHA_0 so $PHA = PHA_0\mathbb{Z} = PHA_0 \times \mathbb{Z}$. Finally, let $T = F \times \{0\} \times \{0\} \leqq E$ so that T is a subgroup of E as well as a transversal for PHA in PHE (and A in E).

Recall that μ_0 denotes the "standard character" of $GF(p) \simeq \mathbb{Z}$, and we may regard μ_0 as a character of $\mathbb{Z}$. Define the representation ρ_1 of PHE by setting $\rho_1 = (1_{PHA_0} \# \mu_0)^{PHE}$ with the understanding that the transversal T is used in the construction of the induced representation. Since T is a group naturally isomorphic to F, $\rho_1(g)$ will be viewed as an $F \times F$ matrix for $g \in PHE$.

All of the above notation will be fixed throughout this section.

The first theorem of this section is an explicit determination of the matrices $\rho_1(g)$ for $g \in PHE$. The result ties in with several of the matrices defined at the end of § 2.

THEOREM 4.1. *The representations ρ_1, $\rho_1|_{PE}$ and $\rho_1|_E$ are all faithful and irreducible. Moreover*

(1) *If $e = (a, b, m) \in E$ then the (r, s) entry of $\rho_1(e)$ is $\delta_{r+a,s}\lambda_{2b}(r + a/2)\mu_0(m)$. In particular, if $e = (0, b, 0) \in A_0$, then $\rho_1(e) = E_{2b}$.*

(2) *If*

$$x = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \in P, \quad \text{then } \rho_1(x) = D_a.$$

(3) *If*

$$h = \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} \in H, \quad \text{then } \rho_1(h) = N_a.$$

Proof. From the definition of ρ_1,

$$\rho_1|_E = (1_{PHA_0} \# \mu_0)^{PHE}|_E = ((1_{PHA_0} \# \mu_0)_A)^E = (1_{A_0} \# \mu_0)^E.$$

The character $1_{A_0} \# \mu_0$ of A has inertia group equal to A in E, and it follows that $(1_{A_0} \# \mu_0)^E$ is an irreducible representation of E. Thus, $\rho_1|_E$ is irreducible as well as $\rho_1|_{PE}$ and ρ_1 itself.

The kernel of $\rho_1|_E$ is the core in E of $\ker(1_{A_0} \# \mu_0) = A_0$. Since every nontrivial normal subgroup of E must intersect the center $\mathbb{Z}$ nontrivially, and $A_0 \cap Z = 1$, it follows that the core of A_0 in E is trivial. Hence $\rho_1|_E$ is faithful. Let K be the kernel of ρ_1. Then $K \cap E = 1$, so $K \subseteq \mathbb{C}_{PHE}(E)$. But $\mathbb{C}_{PHE}(E) \subseteq \mathbb{C}_{PHE}(E/\mathbb{Z}) = E$, as PH acts faithfully on E/Z. Thus $K \subseteq E$ so $K = K \cap E = 1$, and ρ_1 is faithful.

Now suppose $e = (a, b, m) \in E$. For $(r, 0, 0) \in T$ we have $PHA(r, 0, 0)e = PHA(a + r, b, m + \mathrm{tr}(br)) = PHA(a + r, 0, 0)$. Thus, the (r, s) entry of $\rho_1(e)$ is 0 unless $s = r + a$. The $(r, r + a)$ entry is $(1_{PHA_0} \# \mu_0)(x)$, where $x = (a + r, b, m + \mathrm{tr}(br))(-r - a, 0, 0) = (0, b, m + \mathrm{tr}(2br + ba))$. Hence $(1_{PHA_0} \# \mu_0)(x) = \mu_0(m + \mathrm{tr}(2br + ba)) = \mu_0(m)\mu_0(\mathrm{tr}(2b(r + a/2))) = \mu_0(m)\lambda_{2b}(r + a/2)$. The rest of (1)

now follows from the definition of E_{2b}.

Suppose $x = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \in P$. If $(r, 0, 0) \in T$ then $PHA(r, 0, 0)x = PHAxx^{-1}(r, 0, 0)x = PHA(r, 0, 0)^x = PHA((r, 0)x, 0) = PHA(r, ra, 0) = PHA(r, 0, 0)$. It follows that $\rho_1(x)$ is a diagonal matrix, and the (r, r) entry is $(1_{PHA_0} \# \mu_0)(x(r, ra, 0)(-r, 0, 0)) = (1_{PHA_0} \# \mu_0)(0, ra, \operatorname{tr}(r^2 a)) = \mu_0(\operatorname{tr}(r^2 a)) = \lambda_a(r^2)$. From the definition of D_a then, $\rho_1(x) = \rho_1\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} = D_a$.

Finally suppose $h = \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} \in H$. If $(r, 0, 0) \in T$, then $PHA(r, 0, 0)h = PHAhh^{-1}(r, 0, 0)h = PHA(r, 0, 0)^h = PHA((r, 0)h, 0) = PHA(ra, 0, 0)$. Thus, the (r, s) entry of $\rho_1(h)$ is zero unless $s = ra$, and the (r, ra) entry equals

$$(1_{PHA_0} \# \mu_0)(h(ra, 0, 0)(-ra, 0, 0)) = (1_{PHA_0} \# \mu_0)(h) = 1.$$

By definition of N_a then, $\rho_1(h) = \rho_1\begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} = N_a$. □

COROLLARY 4.2. *The group generated by N_a, D_b and E_c for $a, b, c \in F$, $a \neq 0$ is a solvable group of order $(q-1)q^2$.*

Proof. By the preceding theorem,

$$\langle N_a \mid a \in F^\times \rangle = \rho_1(H), \quad \langle D_b \mid b \in F \rangle = \rho_1(P), \quad \langle E_c \mid c \in F \rangle = \rho_1(A_0).$$

It already has been noted that PHA_0 is a group, and hence the group generated by these matrices is $\rho_1(P)\rho_1(H)\rho_1(A_0) = \rho_1(PHA_0)$. As ρ_1 is faithful, this linear group is isomorphic to PHA_0. Clearly PHA_0 has order $(q-1)q^2$ and PHA_0 is contained in the solvable group PHE. □

THEOREM 4.3. *The representation $\rho_1|_{PE}$ is uniquely extendible to a representation ρ of $\tilde{G}$. Moreover, if θ denotes the unique nonprincipal character of PHE/PE satisfying $\theta^2 = 1_{PHE}$, then $\rho|_{PHE} = \theta\rho_1$.*

Proof. First notice that $E \lhd \tilde{G}$ and that the character of $\rho_1|_E$ is invariant in $\tilde{G}$. By Theorem 3.1, to prove that $\rho_1|_E$ is extendible to $\tilde{G}$, it suffices to prove that it is extendible to RE for every Sylow r-subgroup R of S. The case that $r \neq p$ is standard (see [5] for instance). For $r = p$ we may assume $R = P$. In this case, $\rho_1|_{PE}$ is an extension of $\rho_1|_E$, and it follows that $\rho_1|_E$ is extendible to a representation of $\tilde{G}$. Any other extension has the form $\lambda\rho$, where λ is some linear character of $\tilde{G}$.

If $q > 3$ then $\tilde{G}' = \tilde{G}$ forcing $\lambda = 1_{\tilde{G}}$ and so ρ is the unique extension of $\rho_1|_E$ to $\tilde{G}$. Now $\rho|_{PHE}$ and ρ_1 are both extensions of $\rho_1|_E$, so $\rho|_{PHE} = \theta'\rho_1$ for some linear character θ' of PHE. In this case ($q > 3$) the commutator subgroup of PHE is PE, so the kernel of θ' contains PE. The equation $\rho|_{PE} = \rho_1|_{PE}$ follows and ρ is the unique extension of $\rho_1|_{PE}$ to $\tilde{G}$.

Suppose now $q = 3$. Then $|\tilde{G} : \tilde{G}'| = 3$, and there are three extensions of $\rho_1|_E$ to $\tilde{G}$. As before, $\rho|_{PHE} = \theta'\rho_1$ for some linear character θ' of PHE. In this case, however, the commutator subgroup of PHE is E. The character θ' is uniquely determined by its restriction to $PH = P \times H$. Write $\theta' = \theta_1 \# \theta_2$, where θ_1 and θ_2 are irreducible characters of P and H respectively. Let ω denote a nonprincipal linear character of $\tilde{G}$. Since $\tilde{G} = P\tilde{G}'$ it follows that $1_{\tilde{G}}$, ω and ω^2 are the three distinct linear characters of G, and exactly one of these, say ω^i, satisfies $\omega^i|_P = \theta_1$. Then $\omega^{-i}\rho$ is the unique extension of $\rho_1|_E$ to $\tilde{G}$ which satisfies $\omega^{-i}\rho|_{PE} = \rho_1|_{PE}$. We may replace ρ by $\omega^{-i}\rho$ so as to assume $i = 0$ and $\theta_1|_P = 1_P$. Then ρ is the unique extension of $\rho_1|_{PE}$ to $\tilde{G}$.

In either case we have proved that $\rho_1|_{PE}$ has a unique extension to a representation ρ of $\tilde{G}$ and that ρ satisfies $\rho|_{PHE} = \theta'\rho_1$ for some linear character θ' of PHE/PE. It remains to prove that $\theta' = \theta$.

By Theorem 4.1 (3), if $a \in F^\times$ and $h = \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix}$ then $\rho_1(h) = N_a$. Now N_a is a permutation matrix corresponding to the permutation $F \to F$, given by $x \mapsto ax$. Assume a is a generator for $F^\times$. Then this permutation fixes one point (namely 0) and moves the others in an orbit of size $q-1$. As $q-1$ is even, $\det \rho_1(h) = -1$. From the definition of θ, it follows that $\det \rho_1(h) = \theta_H(h)$, holds for all $h \in H$, as the equation holds for a generator.

On the other hand, $\det \rho(h) = 1$ for all $h \in H$ as $H \subseteq \tilde{G}'$ (regardless of the value of q). Thus $1_H = \det \rho|_H = \det(\theta'_H \rho_1|_H) = (\theta'_H)^q \det \rho_1|_H = \theta'_H \theta_H$, as $(\theta'_H)^{q-1} = 1_H$. This equation forces $\theta'_H = \theta_H$, and hence $\theta' = \theta$. The theorem now follows. □

It is worth noting that $\det \rho \neq 1_{\tilde{G}}$, if $q = 3$.

Notice that

$$\theta\begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} = \left(\frac{a}{F}\right),$$

where $(\cdot/F)$ is the "Legendre symbol for F," i.e., $(a/F) = 1$ if a is a square in $F^\times$ and $(a/F) = -1$ if a is a nonsquare in $F^\times$. Thus,

$$\rho\begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} = \left(\frac{a}{F}\right)\rho_1\begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} = \left(\frac{a}{F}\right)N_a \quad \text{for all } a \in F^\times.$$

At this point we have shown that each of the matrices N_a, D_b, E_c (for $a, b, c \in F$, $a \neq 0$) appears in the image of ρ_1. The matrices M_a for $a \neq 0$ are not in the image of ρ_1 but certain scalar multiples of them are in the image of ρ, as the next theorem shows.

THEOREM 4.4. *Let ρ be the unique extension of $\rho_1|_{PE}$ to $\tilde{G}$, guaranteed by Theorem 4.3. Then $\rho\begin{pmatrix} 0 & 1/2 \\ -2 & 0 \end{pmatrix} = cM_1$, where $c = \det M_1$. Moreover, $c = \pm 1$ if $q \equiv 1 \bmod 4$ and $c = \pm i$ if $q \equiv 3 \bmod 4$. Finally,*

$$M_a = c^{-1}\left(\frac{a}{F}\right)\rho\begin{pmatrix} 0 & a/2 \\ -2/a & 0 \end{pmatrix} \quad \text{for all } a \in F^\times.$$

Proof. By Theorem 2.2, $M_1^2 = N_{-1}$ so $M_1^4 = (N_{-1})^2 = I$. In particular, $M_1^{-1} = M_1^3 = N_{-1}M_1$, and the (r, s) entry of M_1^{-1} is $q^{-1/2}\lambda_1(-rs)$. Using this, and Theorem 4.1, the (r, s) entry of $M_1^{-1}\rho_1(a, b, 0)M_1$ is

$$\begin{aligned} &\frac{1}{q}\sum_{t,u \in F} \lambda_1(-rt)\delta_{t+a,u}\lambda_{2b}\left(t + \frac{a}{2}\right)\lambda_1(us) \\ &\quad = \frac{1}{q}\sum_{t \in F} \lambda_1(-rt)\lambda_1(2bt + ab)\lambda_1(ts + as) \\ &\quad = \frac{1}{q}\sum_{t \in F} \lambda_1(ab + as + (-r + 2b + s)t). \end{aligned}$$

If $-r + 2b + s \neq 0$ then the function

$$t \mapsto ab + as + (-r + 2b + s)t,$$

is a bijection of F onto itself, and the corresponding sum is zero. If $-r + 2b + s = 0$,

then each term in the sum equals $\lambda_1(ab+as)=\lambda_a(b+s)$. Thus, the (r, s) entry of $M_1^{-1}\rho_1(a, b, 0)M_1$ is $\delta_{r-2b,s}\lambda_a(b+s)=\delta_{r-2b,s}\lambda_a(r-b)$. From Theorem 4.1 again, this is the (r, s) entry of $\rho_1(-2b, a/2, 0)$. Hence, $M_1^{-1}\rho_1(a, b, 0)M_1=\rho_1(-2b, a/2, 0)$.

Now $x=\begin{pmatrix}0 & 1/2\\ -2 & 0\end{pmatrix}\in S$ and $(-2b, a/2)=(a, b)x$, so that in $\tilde{G}$ $(-2b, a/2, 0)=(a, b, 0)^x=x^{-1}(a, b, 0)x$. Therefore $\rho_1(-2b, a/2, 0)=\rho_1(x^{-1}(a, b, 0)x)=\rho(x^{-1}(a, b, 0)x)=\rho(x)^{-1}\rho(a, b, 0)\rho(x)=\rho(x)^{-1}\rho_1(a, b, 0)\rho(x)$, where we have used $\rho_1|_E=\rho|_E$. This establishes that

$$M_1^{-1}\rho_1(a, b, 0)M_1=\rho\begin{pmatrix}0 & 1/2\\ -2 & 0\end{pmatrix}^{-1}\rho_1(a, b, 0)\rho\begin{pmatrix}0 & 1/2\\ -2 & 0\end{pmatrix}$$

holds for all $a, b\in F$. Hence, $M_1\rho\begin{pmatrix}0 & 1/2\\ -2 & 0\end{pmatrix}^{-1}$ centralizes all matrices of the form $\rho_1(a, b, 0)$. Since $F\times F\times\{0\}$ generates E and $\rho_1|_E$ is irreducible, $M_1\rho\begin{pmatrix}0 & 1/2\\ -2 & 0\end{pmatrix}^{-1}$ is a scalar matrix, by Schur's lemma.

Write $\rho\begin{pmatrix}0 & 1/2\\ -2 & 0\end{pmatrix}=cM_1$. Squaring this equation yields $\rho\begin{pmatrix}-1 & 0\\ 0 & -1\end{pmatrix}=c^2N_{-1}$. But $\begin{pmatrix}-1 & 0\\ 0 & -1\end{pmatrix}\in H$ so $\rho\begin{pmatrix}-1 & 0\\ 0 & -1\end{pmatrix}=(-1/F)\rho_1\begin{pmatrix}-1 & 0\\ 0 & -1\end{pmatrix}=(-1/F)N_{-1}$. Hence, $c^2=(-1/F)$. Now -1 is a square in F if and only if $q\equiv 1 \bmod 4$, and the formulas $c=\pm 1$ for $q\equiv 1 \bmod 4$ and $c=\pm i$ for $q\equiv 3 \bmod 4$ follow. Notice that for all values of q, $c^{q+1}=1$.

The matrix $x=\begin{pmatrix}0 & 1/2\\ -2 & 0\end{pmatrix}$ has order 4 and $|\tilde{G}:\tilde{G}'|$ is odd, so $x\in\tilde{G}'$. Therefore, $\det\rho(x)=1$, which implies $c^q\det M_1=1$ and so $\det M_1=c$.

Finally, notice that for $a\neq 0$, $M_a=M_aM_1^4=(M_aM_1)(M_1^2)M_1=N_{-a}N_{-1}M_1=N_aM_1$. Since $N_a=(a/F)\rho\begin{pmatrix}a & 0\\ 0 & a^{-1}\end{pmatrix}$ and ρ is a homomorphism, we have

$$M_a=\left(\frac{a}{F}\right)\rho\begin{pmatrix}a & 0\\ 0 & a^{-1}\end{pmatrix}c^{-1}\rho\begin{pmatrix}0 & \frac{1}{2}\\ -2 & 0\end{pmatrix}=c^{-1}\left(\frac{a}{F}\right)\rho\begin{pmatrix}0 & \frac{a}{2}\\ -\frac{2}{a} & 0\end{pmatrix},$$

as desired. □

It should be noted that the matrices M_a for $a\in F^\times$ are determined only up to a sign. However, since q is odd, $c=\det M_a$ changes sign when M_a is replaced by $-M_a$. Hence cM_a is uniquely determined.

Because of the presence of scalar factors in the previous theorem, it is convenient to extend the group $\tilde{G}$ slightly. Define $c_q=-1$ if $q\equiv 1 \bmod 4$ and $c_q=i$ if $q\equiv 3 \bmod 4$. Hence $\langle c_q\rangle$ is a cyclic group of order 2 or 4. Define $G_0=\langle c_q\rangle\times S$ and $G=\langle c_q\rangle\times\tilde{G}$. Notice that, since S is naturally embedded in $\tilde{G}$, we may regard G_0 as a subgroup of G.

The representation ρ of $\tilde{G}$ may be extended naturally to G by setting $\rho(c, g)=c\rho(g)$ for $c\in\langle c_q\rangle$ and $g\in\tilde{G}$. Notice that we use the same notation for the extended representation. It is readily checked that ρ is faithful on G.

THEOREM 4.5. *$\langle M_a, N_a, D_b \mid a, b\in F, a\neq 0\rangle=\rho(G_0)$ and $\langle M_a, N_a, D_b, E_b \mid a, b\in F, a\neq 0\rangle=\rho(G)$.*

Proof. Let X denote $\langle M_a, N_a, D_b \mid a, b \in F, a \neq 0\rangle$. Combining all of the previous theorems we have:

$$M_a = c^{-1}\left(\frac{a}{F}\right)\rho\begin{pmatrix} 0 & \frac{a}{2} \\ -\frac{2}{a} & 0 \end{pmatrix}, \quad N_a = \left(\frac{a}{F}\right)\rho\begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix}, \quad D_b = \rho\begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix},$$

where all scalar factors belong to $\langle c_q\rangle$. Hence $X \leqq \rho(G_0)$. Now, the Bruhat decomposition of $SL(2, F) = PH \cup PH\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}P$ is well known, and proves that

$$\left\{\rho\begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}, \rho\begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix}, \rho\begin{pmatrix} 0 & \frac{a}{2} \\ -\frac{2}{a} & 0 \end{pmatrix} \middle| a, b \in F, a \neq 0\right\}$$

generates $\rho(S)$. Hence, modulo $\langle c_qI\rangle$, X is all $\rho(G_0)$, that is $\langle c_qI\rangle X = \rho(G_0)$. It suffices to prove $c_qI \in X$.

Since $\langle c_qI\rangle$ is central in $\langle c_qI\rangle X$, the commutator subgroup of $\langle c_qI\rangle X$ coincides with the commutator subgroup of X. Thus,

$$X' = (\langle c_qI\rangle X)' = \rho(G_0)' = \rho(G_0') = \rho(S').$$

Since $\begin{pmatrix} 0 & a/2 \\ -2/a & 0 \end{pmatrix}$ has order 4 (for any choice of $a \in F^\times$) and $|S:S'|$ is odd, $\rho\begin{pmatrix} 0 & a/2 \\ -2/a & 0 \end{pmatrix}$ belongs to $\rho(S')$ and hence to $X' \subseteq X$. Now $M_a = c^{-1}(a/F)\rho\begin{pmatrix} 0 & a/2 \\ -2/a & 0 \end{pmatrix} \in X$ and so $c^{-1}(a/F)I \in X$. This result holds for all $a \in F$. Since $c^{-1} = \pm c_q$, the field element a may be chosen so that $c^{-1}(a/F) = c_q$. Hence, $c_qI \in X$ and $X = \rho(G_0)$ as required.

Now let $Y = \langle M_a, N_a, D_b, E_b \mid a, b \in F, a \neq 0\rangle$. By Theorem 4.1 (1), $E_b = \rho(0, b/2, 0) \in \rho(\tilde{G}) \leqq \rho(G)$, so $Y \leqq \rho(G)$ by the first part of the proof. Since $X \leqq Y$, the inclusion $\rho(G_0) \leqq Y$ also follows from the first part of the proof.

Since the action of S on $E/\mathbb{Z}$ is irreducible, and $E' = \mathbb{Z}$, it follows that the only subgroups of $\tilde{G}$ containing S are S, $\mathbb{Z}S$ and $\tilde{G}$. Hence, the only subgroups of G containing G_0 are G_0, $\mathbb{Z}G_0$ and G. Now for $b \neq 0$, $E_b \notin \rho(\mathbb{Z}G_0)$ and it follows that $Y = G$ as required. □

The significance of the last result is the fact that the weight enumerator of a self-dual code in F^n is an invariant of $\rho(G_0)$. If the code is also normalized, the weight enumerator is an invariant of $\rho(G)$. Notice that for $q \equiv 3 \bmod 4$ these groups contain iI which forces n to be a multiple of 4. It is possible to give a direct proof of this last result (see [18]).

The exceptional behavior when $q = 3$ is also interesting. Here $\rho(G)$ contains each of the diagonal matrices

$$\begin{pmatrix} \omega & & \\ & 1 & \\ & & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & & \\ & \omega & \\ & & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & & \\ & 1 & \\ & & \omega \end{pmatrix},$$

where ω is a primitive cube root of 1. What this means from a coding theory point of

view is that any code vector from a normalized self-dual code over $GF(3)$ contains a multiple of 3 of zeros, ones and twos.

5. Construction of G (characteristic two). Throughout this section F denotes the field $GF(q)$, where q is a power of 2.

Let $V = F \times F$ and define the bilinear map $\beta: V \times V \to GF(2)$ by $\beta(v, w) = \mathrm{tr}(v_1 w_2)$, where $v = (v_1, v_2)$, $w = (w_1, w_2)$ and $\mathrm{tr}: F \to GF(2)$ is the usual trace map. Associated with β there is another bilinear form β^T (the "transpose of β") given by the familiar formula $\beta^T(v, w) = \beta(w, v)$. If γ is any form and $g \in GL(2, F)$, let $\gamma \cdot g$ denote the form $(v, w) \mapsto \gamma(vg^{-1}, wg^{-1})$. This defines an action $GL(2, F)$ on the space of all bilinear forms $V \times V \to GF(2)$. Set

$$O^+(V) = \{g \in GL(2, F) \mid \beta \cdot g \in \{\beta, \beta^T\}\}.$$

It is readily checked that

$$O^+(V) = \left\{ \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix}, \begin{pmatrix} 0 & b \\ b^{-1} & 0 \end{pmatrix} \middle| a, b \in F^\times \right\},$$

so that $O^+(V)$ is the full two-dimensional orthogonal group over F, with respect to the quadratic form $(v_1, v_2) \mapsto v_1 v_2$.

As in the odd characteristic case, an appropriate extraspecial group E is constructed in such a way that $O^+(V)$ acts "naturally" on E. The matrices M_a, N_a, D_b, E_b ($a \in F^\times$, $b \in F$) will then lie in the image of some irreducible representation of $G = O^+(V) \ltimes E$.

The map β determines the quadratic form $Q(v) = \beta(v, v)$, and notice that β^T determines this same form. If $g \in O^+(V)$, define $\delta(g) \in GF(2)$, by setting $\delta(g) = 0$ if $\beta \cdot g = \beta$, and $\delta(g) = 1$ if $\beta \cdot g = \beta^T$.

The group E is defined as follows. As a set, E is $V \times GF(2)$ (sometimes viewed as $F \times F \times GF(2)$) with multiplication defined by

$$(v, m)(w, n) = (v + w, m + n + \beta(v, w)).$$

It is readily checked that E is an extraspecial 2-group of order $2q^2$. If $g \in O^+(V)$ and $e = (v, m) \in E$ define $e^g = (vg, m + \delta(g)Q(v))$. The equations

$$\delta(gh) = \delta(g) + \delta(h), \quad Q(vg) = Q(v), \quad Q(v + w) = Q(v) + Q(w) + \beta(v, w) + \beta^T(v, w),$$

imply that $O^+(V)$ acts on E by automorphisms.

As usual, let $G = O^+(V) \ltimes E$ be the semidirect product of $O^+(V)$ with E. As in the odd characteristic case, $O^+(V)$ and E are viewed as subgroups of G.

Define four subgroups of E (and hence of G) as follows:

$$A_0 = \{0\} \times F \times \{0\}, \qquad T = F \times \{0\} \times \{0\},$$
$$\mathbb{Z} = \{0\} \times \{0\} \times GF(2), \qquad A = A_0 \mathbb{Z}.$$

Notice that T is a transversal for A in E. Let

$$H = \left\{ \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} \middle| a \in F^\times \right\} \leqq O^+(V).$$

Then H normalizes each of the four subgroups listed above. Also, $HA = HA_0\mathbb{Z} = HA_0 \times \mathbb{Z}$. As in the odd characteristic case, μ_0 (the "standard character" of $GF(2)$) will be viewed as a character of $\mathbb{Z}$. Let ρ_1 denote the representation of HE induced from the character $1_{HA_0} \# \mu_0$ on the subgroup HA. Thus $\rho_1 = (1_{HA_0} \# \mu_0)^{HE}$ and we use the

transversal T in the construction of this induced representation. As before, the rows and columns of $\rho_1(x)$ will be indexed by elements of F.

THEOREM 5.1. *The representations ρ_1 and $\rho_1|_E$ are faithful and irreducible. Moreover*:

(1) *If $e=(a, b, m)\in E$ then the (r, s) entry of $\rho_1(e)$ is $\delta_{r+a,s}\lambda_b(r)\mu_0(m)$. In particular, if $e=(0, b, 0)\in A_0$ then $\rho_1(e)=E_b$.*

(2) *If* $h=\begin{pmatrix} a & 0 \\ 0 & a^{-1}\end{pmatrix}\in H$ *then* $\rho_1(h)=N_a$.

Proof. That ρ_1 and $\rho_1|_E$ are faithful and irreducible follows from an argument identical to that given in the first part of the proof of Theorem 4.1.

Suppose now $e=(a, b, m)\in E$ and $(r, 0, 0)\in T$. Then $HA(r, 0, 0)e = HA(a+r, b, m+\mathrm{tr}\,(rb)) = HA(a+r, 0, 0)$. Hence, the (r, s) entry of $\rho_1(e)$ is zero unless $s=a+r$ while the $(r, r+a)$ entry is $(1_{HA_0}\# \mu_0)(x)$, where $x=(a+r, b, m+\mathrm{tr}\,(rb))(a+r, 0, 0)=(0, b, m+\mathrm{tr}\,(rb))$. Hence $(1_{HA_0}\# \mu_0)(x)=\mu_0(m+\mathrm{tr}\,(rb))=\mu_0(m)\mu_0(\mathrm{tr}\,(rb))=\mu_0(m)\lambda_b(r)$. The rest of (1) follows from the definition of E_b.

Now suppose $a\in F^\times$ and $h=\begin{pmatrix} a & 0 \\ 0 & a^{-1}\end{pmatrix}\in H$. If $(r, 0, 0)\in T$, then $HA(r, 0, 0)h = HAhh^{-1}(r, 0, 0)h = HA(ar, 0, 0)$ so that the (r, s) entry of $\rho_1(h)$ is zero unless $s=ra$. The (r, ra) entry equals $(1_{HA_0}\# \mu_0)(h(ar, 0, 0)(ar, 0, 0))=(1_{HA_0}\# \mu_0)(h)=1$. By the definition of N_a, then, $\rho_1(h)=\rho_1\begin{pmatrix} a & 0 \\ 0 & a^{-1}\end{pmatrix}=N_a$. □

THEOREM 5.2. *The representation ρ_1 of HE is extendible to G. Moreover if ρ denotes one of these extensions, then*

$$\rho\begin{pmatrix} 0 & a \\ a^{-1} & 0\end{pmatrix}=cM_a \quad \text{for } a\in F^\times,$$

where $c=\pm 1$ is independent of a.

Proof. The group E has a unique faithful irreducible character, say ζ, and, by Theorem 5.1, ζ is afforded by $\rho_1|_E$. Let $\hat{\zeta}$ be the character afforded by ρ_1. Then $\hat{\zeta}$ is an extension of ζ to HE and any other extension has the form $\theta\hat{\zeta}$ for some linear character θ of HE/E. Let a be a generator of $F^\times$ and set $h=\begin{pmatrix} a & 0 \\ 0 & a^{-1}\end{pmatrix}\in H$. By the previous theorem, $\rho_1(h)=N_a$. The matrix N_a is a permutation matrix corresponding to the permutation $F\to F$, given by $x\mapsto ax$. There is one fixed point under this permutation (namely 0) and the other elements of F are moved in a cycle of length $q-1$. Since $q-1$ is odd, $\det N_a=1$ and $\det \hat{\zeta}|_H=1_H$ follows.

Suppose first that $q>2$. Then E is the commutator subgroup of HE and $\det\hat{\zeta}=1_{HE}$. Hence $\det\theta\hat{\zeta}=\theta^q\det\hat{\zeta}=\theta$ for any linear character θ of HE/E. This proves that ρ_1 is the unique unimodular representation which extends $\rho_1|_E$, and hence its character $\hat{\zeta}$ is invariant in G. As G/HE is cyclic, ρ_1 is extendible to G, by Theorem 3.1. In fact, as $|G:HE|=2$, there are precisely two extensions of ρ_1 to G.

If $q=2$ then $\rho_1=\rho_1|_E$ and $\zeta=\hat{\zeta}$ is the unique faithful irreducible character of E ($\simeq D_8$). Hence, ρ_1 extends to a representation of G (by Theorem 3.1 again). As in the preceding case, there are two extensions of ρ_1 to G.

Let ρ denote one of the extensions of ρ_1 to G. (The other is then $\kappa\rho$, when κ is the unique linear character of G with kernel HE.)

From Theorem 2.2, $M_1^2 = N_{-1} = N_1 = I$, since $-1 = 1$ in characteristic. 2. Hence $M_1^{-1} = M_1$ and the (r, s) entry of $M_1^{-1}\rho_1(a, b, 0)M_1$ is

$$\frac{1}{q}\sum_{t,u\in F}\lambda_1(rt)\delta_{t+a,u}\lambda_b(t)\lambda_1(us) = \frac{1}{q}\sum_{t\in F}\lambda_1(rt)\lambda_1(bt)\lambda_1(ts+as)$$

$$= \frac{1}{q}\sum_{t\in F}\lambda_1((r+b+s)t+as).$$

Arguing as in the odd characteristic case gives that this sum is zero if $r+b+s \neq 0$, and is $\lambda_1(as) = \lambda_a(s)$ if $r+b+s = 0$. Hence, the (r, s) entry of $M_1^{-1}\rho_1(a, b, 0)M_1$ is $\delta_{r+b,s}\lambda_a(s) = \delta_{r+b,s}\lambda_a(r)\lambda_a(b)$.

On the other hand,

$$\rho\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}^{-1}\rho(a, b, 0)\rho\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} = \rho\left(\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}^{-1}(a, b, 0)\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\right)$$

$$= \rho(b, a, O(a, b)) = \rho(b, a, \operatorname{tr}(ab)).$$

Since $\rho|_E = \rho_1|_E$, the (r, s) entry of this last matrix is $\delta_{r+b,s}\lambda_a(r)\mu_0(\operatorname{tr}(ab)) = \delta_{r+b,s}\lambda_a(r)\lambda_a(b)$. Hence M_1 acts by conjugation on $\rho(E)$ in the same way that $\rho\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$ acts. By Schur's lemma, then, $\rho\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} = cM_1$ for some complex number c. Now $\rho\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$ and M_1 both have order 2, so $c^2 = 1$. Finally, the identity $M_aM_1 = N_{-a} = N_a$ implies $M_a = N_aM_1^{-1} = N_aM_1$. By Theorem 5.1,

$$\rho\begin{pmatrix}a & 0\\ 0 & a^{-1}\end{pmatrix} = \rho_1\begin{pmatrix}a & 0\\ 0 & a^{-1}\end{pmatrix} = N_a,$$

so

$$\rho\begin{pmatrix}0 & a\\ a^{-1} & 0\end{pmatrix} = \rho\left(\begin{pmatrix}a & 0\\ 0 & a^{-1}\end{pmatrix}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\right) = \rho\begin{pmatrix}a & 0\\ 0 & a^{-1}\end{pmatrix}\rho\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} = cN_aM_1 = cM_a,$$

and the theorem follows. □

THEOREM 5.3. *The groups* $\langle M_a, N_a, D_b \mid a \in F^\times, b \in F\rangle$ *and* $\langle M_a, N_a, D_b, E_b \mid a \in F^\times, b \in F\rangle$ *coincide, and both are equal to* $\rho(G)$, *where* ρ *denotes either of the two extensions of* ρ_1 *to* G.

Proof. Let $X = \langle M_a, N_a, D_b \mid a \in F^\times, b \in F\rangle$ and $Y = \langle M_a, N_a, D_b, E_b \mid a \in F^\times, b \in F\rangle$. Obviously $X \leqq Y$. If $b, r \in F$ then $\lambda_{b^2}(r^2) = \lambda(b^2r^2) = \mu_0(\operatorname{tr}(b^2r^2)) = \mu_0(\operatorname{tr}(br)) = \lambda_b(r)$. The penultimate equality follows, since $(br)^2$ is a conjugate of br. Hence, $\lambda_{b^2}(r^2) = \lambda_b(r)$ and $D_{b^2} = E_b$ follows. Thus $X = Y$, proving the first part of the theorem.

Now let ρ denote one of the two extensions of ρ_1 to G and let $c = \pm 1$ be the constant appearing in the statement of Theorem 5.2. Since $\rho(0, 0, 1) = -I$, it follows that

$$M_a = c\rho\begin{pmatrix}0 & a\\ a^{-1} & 0\end{pmatrix} \in \rho(G) \quad \text{for all } a \in F^\times.$$

Theorem 5.1 shows that N_a and $D_b = E_{\sqrt{b}}$ belong to $\rho_1(HE) \leqq \rho(G)$ for all $a \in F^\times$ and $b \in F$. Hence $Y = X \leqq \rho(G)$. It remains to prove the reverse inclusion.

By Theorem 5.1, $\rho(A_0) = \{E_b \mid b \in F\} \leqq Y$. From Theorem 5.2,

$$\rho(T) = \rho\left(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^{-1} A_0 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right) = M_1^{-1}\rho(A_0)M_1 \leqq Y.$$

As T and A_0 generate E, $\rho(E) \leqq Y$. In particular, $-I = \rho(0, 0, 1) \in Y$. Using the same two theorems again, $\rho(O^+(V)) = \langle cM_a, N_a \mid a \in F^\times \rangle \leqq \langle -I, M_a, N_a \mid a \in F^\times \rangle \leqq Y$. Hence, $\rho(G) = \rho(O^+(V))E \leqq Y$, as desired. □

As already noted in the first section, a self-dual code over F is automatically normalized. This fact also follows from the identity $E_b = D_{b^2}$.

The significance of the last theorem is that the complete weight enumerator of a (normalized) self-dual code is an invariant of the linear group $\rho(G)$.

It is worth pointing out that, even though the extension ρ or ρ_1 to G is not unique, the group $\rho(G)$ is uniquely determined. The Molien series for this group is calculated in § 7.

6. The Molien series (odd characteristic). The Molien series associated with a complex representation ρ of a finite group G is defined by

$$\Phi_{G,\rho}(X) = \frac{1}{|G|} \sum_{g \in G} \frac{1}{\det(I - X\rho(g))}.$$

The coefficient of X^d in this expansion is the dimension of the space of homogeneous polynomials of degree d which are invariant under the action of the matrix group $\rho(G)$ [17]. A proof of this fact may also be found in [19]. When the representation ρ of G is clear from the context, the Molien series will be denoted by $\Phi_G(X)$. Notice that the polynomial $\det(I - X\rho(g))$ depends only on the eigenvalues of $\rho(g)$ and hence on the conjugacy class of g, rather than on g itself. In calculating the series it is convenient to group together terms from the same conjugacy class. The Molien series is determined, then, when a complete description is given for the conjugacy classes, and when for each class representative g the polynomial $\det(I - X\rho(g))$ is known.

It is convenient to begin this section with two lemmas concerning quadratic forms over finite fields. These lemmas will be useful in calculating some of the polynomials $\det(I - X\rho(g))$.

LEMMA 6.1. *Let F be a finite field of odd characteristic and let F_0 be the prime subfield of F. If $0 \neq c \in F$, let $Q_c : F \to F_0$ denote the quadratic form given by $Q_c(x) = \operatorname{tr}(cx^2)$, where $\operatorname{tr} : F \to F_0$ is the trace map. If $a \in F$ is not a square, then Q_a is not equivalent to Q_1.*

Proof. If $d \in F$, $d \neq 0$, then Q_c is equivalent to Q_{cd^2} (the transforming map $F \to F$ is given by multiplication by d). Hence, we may assume that a is a generator for the multiplicative group $F^\times$. If $|F| = p^n$, where p is a prime, then $1, a, a^2, \cdots, a^{n-1}$ is a basis for F/F_0. Let $u_0, u_1, \cdots, u_{n-1}$ be the dual basis. (That is, $\operatorname{tr}(a^i u_j) = \delta_{ij}$.) Define the matrix $M = (m_{ij})$ over F_0 by the equations

$$a^i = \sum_{j=0}^{n-1} m_{ij}u_j, \qquad i = 0, 1, \cdots, n-1.$$

Hence, $\operatorname{tr}(a^i a^j) = m_{ij} = m_{ji}$. Notice that M is the matrix of the form Q_1 in the basis $1, a, \cdots, a^{n-1}$. Let $f(x) = x^n + r_{n-1}x^{n-1} + \cdots + r_0 \in F_0[x]$ be the irreducible polynomial of a over F_0. Then the matrix of the form Q_a in this same basis is easily worked out to be $CM^T = CM$, where C is the companion matrix of $f(x)$. Therefore, the discriminant of Q_1 is $\det M$, and that of Q_a is $(\det M)(\det C) = (-1)^n r_0 \det M$. To prove that Q_1 is not equivalent to Q_a, it suffices to show that $(-1)^n r_0$ is not a square in F_0. The norm

map $N: F^\times \to F_0^\times$ is surjective, and the generator a of $F^\times$ must map to a generator of $F_0^\times$. Hence, $N(a) = (-1)^n r_0$ generates F_0, and in particular cannot be a square in F_0. The lemma now follows. □

LEMMA 6.2. *Let p be an odd prime and $l \in GF(p)^\times$ a nonsquare. Let Q_d^+ and Q_d^- be the quadratic forms defined over $GF(p)$ by the equations*:

$$Q_{2k+1}^+(x_1, x_2, \cdots, x_{2k+1}) = x_1x_2 + \cdots + x_{2k-1}x_{2k} + x_{2k+1}^2, \qquad k \geqq 0,$$

$$Q_{2k+1}^-(x_1, x_2, \cdots, x_{2k+1}) = x_1x_2 + \cdots + x_{2k-1}x_{2k} + lx_{2k+1}^2, \qquad k \geqq 0,$$

$$Q_{2k}^+(x_1, x_2, \cdots, x_{2k}) = x_1x_2 + \cdots + x_{2k-1}x_{2k}, \qquad k \geqq 1,$$

$$Q_{2k}^-(x_1, x_2, \cdots, x_{2k}) = x_1x_2 + \cdots + x_{2k-3}x_{2k-2} + x_{2k-1}^2 - lx_{2k}^2, \qquad k \geqq 1.$$

For $c \in GF(p)$ let $n_d^\pm(c)$ denote $|\{(x_1, x_2, \cdots, x_d) \in GF(p)^d \mid Q_d^\pm(x_1, \cdots, x_d) = c\}|$. Then:

(1) $n_{2k+1}^+(0) = n_{2k+1}^-(0) = p^{2k}$, $n_{2k}^+(0) = p^{2k-1} + p^k - p^{k-1}$, $n_{2k}^-(0) = p^{2k-1} - p^k + p^{k-1}$;

(2) $n_{2k}^+(1) = n_{2k}^+(l) = p^{2k-1} - p^{k-1}$;

(3) $n_{2k+1}^+(1) = n_{2k+1}^-(l) = p^{2k} + p^k$, $n_{2k+1}^+(l) = n_{2k+1}^-(1) = p^{2k} - p^k$;

(4) $n_{2k}^-(1) = n_{2k}^-(l) = p^{2k-1} + p^{k-1}$.

The proof follows from straightforward counting arguments and will be omitted. In verifying the fourth identity, it is useful to observe $Q_{2k}^- = Q_{2k-2}^+ \dot{+} Q_2^-$, and that Q_2^- is equivalent to the quadratic form $GF(p^2) \to GF(p)$, given by the norm map. All of the notation from § 4 is retained in this section. In particular, for the remainder of this section, F is the field $GF(q)$, where q is a power of the odd prime p. The groups $\tilde{G} = SL(2, F) \ltimes E$, $S = SL(2, F)$ are direct factors of G, G_0, respectively. Because of this, the Molien series for G and G_0 with respect to ρ follow easily from those of $\tilde{G}$ and S (see Theorems 6.9 and 6.11 below).

The Molien series for S will be calculated first. Let ψ be the character afforded by the representation ρ. Then ψ is irreducible since, in fact, ψ_E is irreducible. However, $\rho\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = (-1/F)N_{-1}$ is not a scalar matrix, so ψ_S reduces. If $(-1/F) = 1$, then $\rho\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ has eigenvalues 1 and -1 with multiplicity $(q+1)/2$ and $(q-1)/2$, respectively, while if $(-1/F) = -1$, these multiplicities are reversed. Thus, ψ_S reduces to a sum of two characters, say $\xi + \eta$, where the degrees are $(q \pm 1)/2$ and where no constituent of one of the summands is faithful while all of the constituents of the other are.

In fact, the characters ξ, η are irreducible, as is seen by the following. By standard arguments $\psi\bar{\psi}$ is the character of the representation of G on $\mathcal{M} = \mathcal{M}at_{q\times q}(\mathbb{C})$ where the action is given by $m \cdot g = \rho(g)^{-1} m \rho(g)$ for $g \in G$ and $m \in \mathcal{M}$. Now $\rho(F \times F \times \{0\})$ is a basis for $\mathcal{M}$, and S permutes this basis. Thus, $\psi_S\bar{\psi}_S$ is the permutation character of S in its action on $F \times F \times \{0\}$. There are exactly two orbits (one contains only $(0, 0, 0)$) so the multiplicity of 1_S in $\bar{\psi}_S\psi_S$ is 2. Hence, by properties of character inner products,

$$(\psi_S, \psi_S)_S = (\psi_S\bar{\psi}_S, 1_S)_S = 2,$$

proving that ξ and η are irreducible.

The decomposition $\psi_S = \xi + \eta$ can also be seen by appealing directly to the character table (see [4, p. 228]). Indeed, by Theorems 4.1 and 4.3, we know that $\rho\begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} = (a/F)N_a$ holds for all $a \in F^\times$. Since N_a is a permutation matrix correspond-

ing to the permutation $F \to F$ given by multiplication by a, it follows that $\psi_H = \theta_H +$ (regular character of H). Here θ has the same meaning as in Theorem 4.3. In particular, 1_S appears at most once as a constituent of ψ_S. Any other irreducible character of $SL(2, F)$ of degree less than q has degree $q-1$ or $(q \pm 1)/2$. The restriction of any character of degree $q-1$ to H is the regular character of H, and as such contains 1_H as a constituent. If this character appears in ψ_S then so must 1_S and this leads to the contradiction $(\psi_H, 1_H)_H = 2$. From the character table of $SL(2, F)$ it now follows that $\psi_S = \xi_i + \eta_i$, where i is 1 or 2, using the notation of the table found in Dornhoff's book.

Using the ideas appearing in [12], the values of the character ψ_S can be computed explicitly without the use of a character table, although we shall not do this here. Notice that the eigenvalues of $\rho(g)$ are readily computable if g is in P or H since these matrices are known and easy to work with. The character ψ_S will be used to determine the polynomials $\det(I - X\rho(g))$ for other elements of S.

Let a be a generator of $F^\times$ and set $h = \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix}$. Let $k \in SL(2, 4)$ have order $q+1$ and set $c = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $d = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}$. Let 1 and z denote the identity matrix and $\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$, respectively. Table 1 lists the conjugacy classes of $SL(2, F)$ together with the size of each class. This differs from Dornhoff's table [4, p. 288] only in that he uses a and b to denote h and k.

TABLE 1.

Class representative	1	z	h^l, $1 \leqq l \leqq (q-3)/2$	k^m, $1 \leqq m \leqq (q-1)/2$	c	d	cz	dz
Size of class	1	1	q^2+q	q^2-q	$(q^2-1)/2$	$(q^2-1)/2$	$(q^2-1)/2$	$(q^2-1)/2$

As in Dornhoff's book, it is convenient to let $\varepsilon = (-1/F)$. Thus $\varepsilon = 1$ if $q \equiv 1 \bmod 4$ and $\varepsilon = -1$ if $q \equiv 3 \bmod 4$. To calculate $\det(I - X\rho(g))$ we consider each class in turn.

Let $g = h^l \in H$. We already observed that

$$\psi_H = \theta_H + (\text{regular character of } H).$$

Since $\langle h^l \rangle$ has index $d = (l, q-1)$ in H we have $\psi_{\langle h^l \rangle} = \theta|_{\langle h^l \rangle} + d$ (regular character of $\langle h^l \rangle$). As $\theta(h^l) = (-1)^l = (-1)^d$ it now follows that $\det(I - X\rho(h^l)) = (1-(-1)^d X)(1 - X^{(q-1)/d})^d$. Notice that this polynomial depends only on $d = (l, q-1)$ rather than l, and this will be used later. Now suppose $K = \langle k \rangle$ and $g = k^m \in K$. Since $\psi_S = \xi_i + \eta_i$, where i is 1 or 2, it follows from the character table that

$$(\text{regular character of } K) = \psi_K + \mu,$$

where μ is the unique nonprincipal character of K satisfying $\mu^2 = 1_K$, i.e., $|K : \ker \mu| = 2$. If $d = (m, q+1)$ then, arguing as in the previous case, we have $\det(I - X\rho(k^m)) = (1 - X^{(q+1)/d})^d / (1 - (-1)^d X)$. As before, this polynomial depends only on $d = (m, q+1)$ rather than on m.

Notice that $z = h^{(q-1)/2} = k^{(q+1)/2}$ so that $\det(I - X\rho(z))$ may be computed using either of the two formulas derived, alone. This polynomial equals $(1-X^2)^{(q-1)/2}(1 - \varepsilon X) = (1-X^2)^{(q+1)/2}/(1+\varepsilon X)$, where as we recall $\varepsilon = (-1)^{(q-1)/2} = (-1/F)$. Obviously, $\det(I - X\rho(1)) = (1-X)^q$. It remains to calculate $\det(I - X\rho(g))$ for the elements of orders divisible by p.

Since $\rho\begin{pmatrix}1 & b\\ 0 & 1\end{pmatrix} = D_b$ for all $b \in F$, and each D_b is a diagonal matrix, it readily follows that

$$\det(I - X\rho(c)) = \prod_{r\in F}(1-\lambda_1(r^2)X)$$

and

$$\det(I - X\rho(d)) = \prod_{r\in F}(1-\lambda_a(r^2)X).$$

Denote these two polynomials by $f_+(X)$ and $f_-(X)$ respectively. Finally, since $\rho(z) = \varepsilon N_{-1}$, we have $\rho(cz) = \varepsilon D_1 N_{-1}$. This matrix is similar to a matrix in blocked diagonal form which contains a single 1×1 block consisting of ε (corresponding to $r = 0$) and $(q-1)/2$ blocks of size 2×2 of the form

$$\begin{pmatrix}0 & \varepsilon\lambda(r^2)\\ \varepsilon\lambda(r^2) & 0\end{pmatrix}$$

(corresponding to the unordered pair $\{r, -r\}$ for $r \in F^\times$). Thus, $\det(I - X\rho(cz)) = (1-\varepsilon X)\prod(1-\lambda(r^2)^2X^2)$, where the product extends over a transversal for $\{1, -1\}$ in $F^\times$. Let this polynomial be denoted by $g_+(X)$. Replacing cz by dz amounts to changing λ_1 to λ_a in the formula. If the resulting polynomial is denoted by $g_-(X)$, then $g_-(X) = \det(I - X\rho(dz)) = (1-\varepsilon X)\prod(1-\lambda_a(r^2)^2X^2)$, where the product extends over the same index set as before.

The only obstacle to the calculation of the Molien series for $S = SL(2, F)$ (and hence G_0) is the determination of the polynomials $f_\pm(X)$ and $g_\pm(X)$. Notice that $f_+(X)$ and $f_-(X)$ enter symmetrically into the Molien series (as do $g_+(X)$ and $g_-(X)$) so only the unordered pairs $\{f_+(X), f_-(X)\}$ and $\{g_+(X), g_-(X)\}$ need be determined.

LEMMA 6.3. $f_+(X)f_-(X) = (1-X^p)^{2q/p}$.

Proof. From the definition of $f_\pm(X)$,

$$f_\pm(X) = \prod_{r\in F}(1-\lambda(cr^2)X),$$

where the subscript is $+$ if $c \neq 0$ is a square, and $-$ if c is a nonsquare. Clearly,

$$f_+(X)f_-(X) = (1-X)^2\Big[\prod_{r\in S}(1-\lambda(r)X)\Big]^2\Big[\prod_{r\in N}(1-\lambda(r)X)\Big]^2,$$

where S and N denote the set of squares and nonsquares in $F^\times$. Hence,

$$f_+(X)f_-(X) = \Big[\prod_{r\in F}(1-\lambda(r)X)\Big]^2.$$

Now the map $r \mapsto \lambda(r)$ is a q/p to one map from F onto the group of pth roots of 1, so $f_+(X)f_-(X) = (1-X^p)^{2q/p}$ as desired. □

Let Q_p denote the field generated over Q by adjoining a primitive pth root of unity. Notice that $\lambda(r) \in Q_p$ for all $r \in F$, so $f_+(X)$ and $f_-(X)$ have coefficients in Q_p. More is true, as the following lemma shows.

LEMMA 6.4. *If q is a square, then $f_+(X)$ and $f_-(X)$ belong to $\mathbb{Z}[X]$. If q is not a square, then $f_+(X)$ and $f_-(X)$ are conjugate polynomials having coefficients in the unique quadratic extension of Q contained in Q_p.*

Proof. Let $\alpha \in \mathrm{Gal}(Q_p/Q)$. Then α sends a primitive pth root of unity to its lth power for some $l \in \mathbb{Z}$ with $(l, p) = 1$. Hence $\lambda(r)^\alpha = \lambda(r)^l = \lambda(lr)$ for all $r \in F$. In the last

expression, l is regarded as a nonzero element of $GF(p)$. Hence

$$f_+(X)^\alpha = \prod_{r\in F}(1-\lambda(r^2)^l X) = \prod_{r\in F}(1-\lambda(lr^2)X).$$

If l is a square in F, this last expression is $f_+(X)$. Otherwise it is $f_-(X)$.

If q is a square, then every element of $GF(p)$ is a square in F, so $f_+(X)^\alpha = f_+(X)$, holds for all $\alpha \in \mathrm{Gal}(Q_p/Q)$. Hence $f_+(X)$, and similarly $f_-(X)$, have coefficients in $\mathbb{Z}$.

If q is not a square, then α may be chosen so that l is not a square in F. Thus $\{f_+(X), f_-(X)\}$ is an orbit under the action of $\mathrm{Gal}(Q_p/Q)$ and this implies the final assertion of the lemma. □

LEMMA 6.5. *Assume q is a square. Then, in some order $f_+(X)$ and $f_-(X)$ are the polynomials*

$$(1-X)^{q/p-\sqrt{q}+\sqrt{q}/p}(1+X+\cdots+X^{p-1})^{q/p+\sqrt{q}/p}$$

and

$$(1-X)^{q/p+\sqrt{q}-\sqrt{q}/p}(1+X+\cdots+X^{p-1})^{q/p-\sqrt{q}/p}.$$

Proof. By definition, $f_+(X) = \prod_{r\in F}(1-\lambda(r^2)X)$. By the two preceding lemmas, and the irreducibility of the cyclotomic polynomial, $f_+(X)$ has the form $(1-X)^l(1+X+\cdots+X^{p-1})^m$, where, of course, $l+(p-1)m = q$. In particular, $f_+(X)$ is completely determined by the multiplicity of the root 1. This multiplicity is $|\{r\in F|\lambda(r^2)=1\}|$. Now, $\lambda(r^2)=\mu_0(\mathrm{tr}(r^2))=1$ if and only if $\mathrm{tr}(r^2)=0$, so $l=|\{r\in F|\,\mathrm{tr}(r^2)=0\}|$. Similarly, the multiplicity of the root 1 for $f_-(X)$ is $|\{r\in F|\,\mathrm{tr}(ar^2)=0\}|$ where $a\in F^\times$ is a nonsquare. By Lemma 6.1, the forms $x\mapsto \mathrm{tr}(x^2)$ and $x\mapsto \mathrm{tr}(ax^2)$ are not equivalent. Write $q=p^{2k}$, so that F is isomorphic to the vector space $GF(p)^{2k}$, and these two forms are regarded as forms defined on $GF(p)^{2k}$. Now in any given dimension, there are two equivalence classes of nonsingular quadratic forms. Hence these two forms (in some order) are equivalent to Q^+_{2k} and Q^-_{2k}, using the notation of Lemma 6.2. By that lemma, l is either $p^{2k-1}+p^k-p^{k-1}$ or $p^{2k-1}-p^k+p^{k-1}$, and Lemma 6.5 follows. □

It is possible to determine explicitly which polynomial is $f_+(X)$, although this will not be needed. The result is that $f_+(X)$ is the first polynomial if $p\equiv 1 \bmod 4$, and is the second polynomial if $p\equiv 3 \bmod 4$.

LEMMA 6.6. *Assume $q=p^{2k+1}$ is not a square and that $\Phi_1(X)$ and $\Phi_2(X)$ are the irreducible factors of $(1+X+\cdots+X^{p-1})$ over the quadratic extension of Q contained in Q_p. Assume also that Φ_1 and Φ_2 are normalized so that $\Phi_1(0)=\Phi_2(0)=1$. Then, in some order, $f_+(X)$ and $f_-(X)$ are the polynomials*

$$(1-X)^{q/p}\Phi_1(X)^{p^{2k}+p^k}\Phi_2(X)^{p^{2k}-p^k}$$

and

$$(1-X)^{q/p}\Phi_1(X)^{p^{2k}-p^k}\Phi_2(X)^{p^{2k}+p^k}.$$

Proof. By Lemmas 6.3 and 6.4, $f_+(X)$ has the form $(1-X)^{q/p}\Phi_1(X)^l\Phi_2(X)^m$, where $q/p+(p-1)(l+m)/2=q$ and $f_-(X)$ has the same form with l and m interchanged. The lemma will follow once l is computed. Since $\mu_0(1)$ is a primitive pth root of 1, choose the notation so that $\mu_0(1)$ is a root of $\Phi_1(X)$. Then the multiplicity l of $\Phi_1(X)$ as a factor of $f_+(X)$ is the same as for the factor $(1-\mu_0(1)X)$. Now $\lambda(r^2)=\mu_0(\mathrm{tr}(r^2))$ so $\lambda(r^2)=\mu_0(1)$ if and only if $\mathrm{tr}(r^2)=1$. Since $f_+(X)=\prod_{r\in F}(1-\lambda(r^2)X)$, we have $l=|\{r\in F|\,\mathrm{tr}(r^2)=1\}|$.

From the same argument as in the previous lemma, this integer is $n^+_{2k+1}(1)=p^{2k}+p^k$, or $n^-_{2k+1}(1)=p^{2k}-p^k$ (in the notation of Lemma 6.2), and the result follows. □

Notice that when q is not a square, $f_+(X) \neq f_-(X)$, so $f_+(X)$ does not have all coefficients in $\mathbb{Z}$. It is worth noting that when $q = p$ is a prime, then these polynomials are $(1-X)\Phi_1(X)^2$ and $(1-X)\Phi_2(X)^2$.

Fortunately, the polynomials $g_\pm(X)$ can be calculated in terms of $f_\pm(X)$, as the next lemma shows. For convenience, if $c \in F^\times$ define $f_c(X) = f_+(X)$ and $g_c(X) = g_+(X)$ if c is a square, and $f_c(X) = f_-(X)$, $g_c(X) = g_-(X)$ otherwise. Recall that $\varepsilon = (-1/F) = (-1)^{(q-1)/2}$.

LEMMA 6.7. *If $c \in F^\times$ then*

$$g_c(X) = (1-\varepsilon X)\left[\frac{f_{2c}(X^2)}{1-X^2}\right]^{1/2},$$

where the square root is chosen so that $g_a(0) = 1$.

Proof. By definition,

$$g_c(X) = (1-\varepsilon X)\prod_{r\in\Lambda}(1-\lambda_c(r^2)^2X^2),$$

where Λ is a set of coset representatives for $\{\pm 1\}$ in $F^\times$. Hence,

$$\begin{aligned} g_c(X) &= (1-\varepsilon X)\prod_{r\in\Lambda}(1-\lambda_{2c}(r^2)X^2) \\ &= (1-\varepsilon X)\left[\prod_{r\in F^\times}(1-\lambda_{2c}(r^2)X^2)\right]^{1/2}. \end{aligned}$$

The product inside the square root is $f_{2c}(X^2)$ except that the term corresponding to $r = 0$ is not present. Since this term is $(1-X^2)$, the lemma follows. □

As a consequence of this lemma and Lemma 6.3 we have

$$g_+(X)g_-(X) = (1-\varepsilon X)^2\frac{(1-X^{2p})^{q/p}}{1-X^2}.$$

Notice also that when $q = p$ is a prime, $g_\pm(X)$ are equal to $(1-\varepsilon X)\Phi_1(X^2)$ and $(1-\varepsilon X)\Phi_2(X^2)$.

At this point, the polynomials $\det(I - X\rho(g))$ have been determined for all $g \in S = SL(2, F)$. The conjugacy classes for this group have already been listed, and this leads easily to the next result.

THEOREM 6.8. *The Molien series for S with respect to ρ may be calculated as follows*:

$$\begin{aligned} (q^3-q)\Phi_S(X) = {} & \frac{1}{(1-X)^q} + \frac{1}{(1-\varepsilon X)(1-X^2)^{(q-1)/2}} \\ & + \frac{q(q+1)}{2}\sum_{\substack{d\mid q-1 \\ d\neq(q-1)/2,\,q-1}} \frac{\phi\left(\frac{q-1}{d}\right)}{(1-X^{(q-1)/d})^d(1-(-1)^dX)} \\ & + \frac{q(q-1)}{2}\sum_{\substack{d\mid q+1 \\ d\neq(q+1)/2,\,q+1}} \phi\left(\frac{q+1}{d}\right)\frac{(1-(-1)^dX)}{(1-X^{(q+1)/d})^d} \\ & + \frac{q^2-1}{2}\left\{\left[\frac{f_+(X)+f_-(X)}{(1-X^p)^{2q/p}}\right]\right. \\ & \left. + \left[(g_+(X)+g_-(X))\frac{(1-X^2)}{(1-\varepsilon X)^2(1-X^{2p})^{q/p}}\right]\right\}. \end{aligned}$$

Here $f_{\pm}(X)$ is given by Lemmas 6.5 or 6.6 according to whether or not q is a square, and $g_{\pm}(X)$ is given by Lemma 6.7.

Recall that $G_0 = \langle c_q \rangle \times S$, where $\langle c_q \rangle$ is -1 or i, according as $q \equiv 1 \bmod 4$ or $q \equiv 3 \bmod 4$.

THEOREM 6.9. *The Molien series for G_0 with respect to ρ may be calculated as follows: If $q \equiv 1 \bmod 4$ then*

$$\Phi_{G_0}(X) = \tfrac{1}{2}(\Phi_S(X) + \Phi_S(-X)),$$

while if $q \equiv 3 \bmod 4$ then

$$\Phi_{G_0}(X) = \tfrac{1}{4}(\Phi_S(X) + \Phi_S(-X) + \Phi_S(iX) + \Phi_S(-iX)).$$

In either case, $\Phi_S(X)$ may be calculated as in Theorem 6.8.

Proof. As $\Phi_{G_0}(X) = 1/|G_0| \sum_{g \in G_0} 1/\det(I - X\rho(g))$ and $G_0 = \langle c_q \rangle \times S$ the contribution to the sum coming from the coset $c_q^i S$ is

$$\frac{1}{|G_0|} \sum_{g \in S} \frac{1}{\det(I - c_q^i X \rho(g))} = \left(\frac{1}{|\langle c_q \rangle|}\right) \Phi_S(c_q^i X).$$

The result follows by summing over j, $0 \leq j < |\langle c_q \rangle|$. □

Much of the work has already been done towards the calculation of the Molien series for $\tilde{G} = SL(2, F) \ltimes E$ and $G = \langle c_q \rangle \times \tilde{G}$. Table 2 is a list of the conjugacy classes of $\tilde{G}$. Here h and $k \in S$ have the same meaning as before. Again, Λ denotes a set of coset representatives for $\{\pm 1\}$ in $F^{\times}$.

Let $\mathbb{Z} = \mathbb{Z}(\tilde{G}) = \mathbb{Z}(E)$ so that $\mathbb{Z}$ is cyclic of order p and generated by $\zeta = (0, 0, 1) \in E$. If $\mathscr{C}$ is a $\tilde{G}$-conjugacy class then so is $\mathscr{C}\zeta^j$ and hence $\mathbb{Z}$ acts on the set of conjugacy classes by multiplication. Since $|\mathbb{Z}| = p$, all orbits have size 1 or p. If a representative is listed in the first column as $g\zeta^j$ then the corresponding class is in an orbit of size p under the action of $\mathbb{Z}$, and it is understood that j ranges over the integers from 0 to $p-1$. If the representative is not listed in this form, then the corresponding class is fixed by $\mathbb{Z}$. These classes appear in the last three lines of the table.

TABLE 2
Conjugacy classes of $\tilde{G}$.

Representative	Number of classes	Size of class
ζ^j	p	1
$z\zeta^j$	p	q^2
$h^i\zeta^j$, $1 \leq i \leq (q-3)/2$	$p(q-3)/2$	$q^3(q+1)$
$k^i\zeta^j$, $1 \leq i \leq (q-1)/2$	$p(q-1)/2$	$q^3(q-1)$
$c\zeta^j$	p	$q(q^2-1)/2$
$d\zeta^j$	p	$q(q^2-1)/2$
$cz\zeta^j$	p	$q^2(q^2-1)/2$
$dz\zeta^j$	p	$q^2(q^2-1)/2$
$(1, 0, 0) \in E - \mathbb{Z}$	1	$q^2p - p$
$c(r, 0, 0)$, $r \in \Lambda$	$(q-1)/2$	$pq(q^2-1)$
$d(r, 0, 0)$, $r \in \Lambda$	$(q-1)/2$	$pq(q^2-1)$

Notice that a given row may correspond to several classes. The number of such classes is given by the entry in the second column, and the (common) size of any of the classes is given in the third column.

It can be checked that these representatives do in fact lie in distinct classes, and the fact that the dot product of the last two columns equals $pq^3(q^2-1)=|\tilde{G}|$ shows that the list is complete.

The first eight lines correspond to class representatives of the form $g\zeta^i$, where $g \in S$. Moreover, if $f(X)=\det(I-\rho(g)X)$ then $\det(I-X\rho(g\zeta^i))=\det(I-\delta^iX\rho(g))=f(\delta^iX)$, where $\delta=\mu_0(1)$ is a primitive pth root of 1. These polynomials have already been calculated, and so it remains to determine $\det(I-X\rho(g))$ for g being a class representative appearing in one of the last three lines.

If $p>3$ then $g^p=1$ holds for each of these representatives. Moreover, g is conjugate to $g\zeta$ which implies that multiplication by δ (a primitive pth root of 1) is a permutation of the eigenvalues of $\rho(g)$, preserving multiplicity. Hence, each eigenvalue δ^i appears with the same multiplicity q/p in $\rho(g)$ and it follows that

$$\det(I-X\rho(g))=\left[\prod_{j=0}^{p-1}(1-\delta^jX)\right]^{q/p}=(1-X^p)^{q/p}.$$

If $p=3$ and $g=(1,0,0)$, the same argument as above applies, and yields $\det(I-X\rho(g))=(1-X^3)^{q/3}$. However, the elements $c(r,0,0)$ and $d(r,0,0)$ may have order 9 (depending on the value of r), and a different argument is necessary to calculate $\det(I-X\rho(g))$.

Assume then $p=3$. Direct calculation shows $(c(r,0,0))^3=(0,0,\operatorname{tr}(-r^2))$ and $(d(r,0,0))^3=(0,0,\operatorname{tr}(-ar^2))$, where $a\in F^\times$ is a nonsquare. Let $\omega=\mu_0(1)$ so that ω is a cube root of 1. Then $\rho((c(r,0,0))^3)=\omega^{\operatorname{tr}(-r^2)}I$ and $\rho((d(r,0,0))^3)=\omega^{\operatorname{tr}(-ar^2)}I$. Therefore, if g denotes $c(r,0,0)$ (or $d(r,0,0)$) then the eigenvalues of $\rho(g)$ are cube roots of $\omega^{\operatorname{tr}(-r^2)}$ (or $\omega^{\operatorname{tr}(-ar^2)}$). Since g is conjugate to $g\zeta$, multiplication by ω permutes the eigenvalues of $\rho(g)$, preserving multiplicity. Hence, if δ is an eigenvalue of g, then $\delta^3=\omega^{\operatorname{tr}(-r^2)}$ (or $\omega^{\operatorname{tr}(-ar^2)}$) and

$$\begin{aligned}\det(I-X\rho(g))&=[(1-\delta X)(1-\delta\omega X)(1-\delta\omega^2X)]^{q/3}\\&=(1-\delta^3X^3)^{q/3}.\end{aligned}$$

This last expression is $(1-\omega^{\operatorname{tr}(-r^2)}X^3)^{q/3}$ for $g=c(r,0,0)$ and $(1-\omega^{\operatorname{tr}(-ar^2)}X^3)^{q/3}$ for $g=d(r,0,0)$.

The contribution to the Molien series for $\tilde{G}$ from all the conjugacy classes represented in the last two lines of the table is

$$\begin{aligned}3q(q^2-1)&\sum_{r\in\Lambda}\left(\frac{1}{(1-\omega^{\operatorname{tr}(-r^2)}X^3)^{q/3}}+\frac{1}{(1-\omega^{\operatorname{tr}(-ar^2)}X^3)^{q/3}}\right)\\&=\frac{3q(q^2-1)}{2}\sum_{r\in F}\left(\frac{1}{(1-\omega^{\operatorname{tr}(-r^2)}X^3)^{q/3}}+\frac{1}{(1-\omega^{\operatorname{tr}(-ar^2)}X^3)^{q/3}}\right)\\&\quad-\frac{3q(q^2-1)}{2}\cdot\frac{2}{(1-X^3)^{q/3}}.\end{aligned}$$

If $q=3^m$, then the forms $x\mapsto\operatorname{tr}(-x^2)$ and $x\mapsto\operatorname{tr}(-ax^2)$ may be identified with Q_m^+ and Q_m^- in some order. By Lemma 6.2, then, these sums reduce to

$$\begin{aligned}&\frac{3q(q^2-1)}{2}\left\{\frac{n_m^+(0)+n_m^-(0)-2}{(1-X^3)^{q/3}}+\frac{n_m^+(1)+n_m^-(1)}{(1-\omega X^3)^{q/3}}+\frac{n_m^+(-1)+n_m^-(-1)}{(1-\omega^2X^3)^{q/3}}\right\}\\&\quad=q^2(q^2-1)((1-X^3)^{-q/3}+(1-\omega X^3)^{-q/3}+(1-\omega^2X^3)^{-q/3})\\&\qquad-3q(q^2-1)(1-X^3)^{-q/3}.\end{aligned}$$

This may be combined, with the contribution to the Molien series by the class represented by (1, 0, 0), to yield

$$(*)\ S(X)=(q^2-3q+3)(q^2-1)(1-X^3)^{-q/3}+q^2(q^2-1)((1-\omega X^3)^{-q/3}+(1-\omega^2X^3)^{-q/3}).$$

Gathering together all the information obtained thus far, we have the following theorem.

THEOREM 6.10. *The Molien series for $\tilde{G}$ with respect to ρ is given by*

$$pq^3(q^2-1)\Phi_{\tilde{G}}(X)=\sum_{j=0}^{p-1}\Phi(\delta^jX)+S(X),$$

where δ is a primitive pth root of 1. Here

$$\Phi(X)=\frac{1}{(1-X)^q}+\frac{q^2}{(1-\varepsilon X)(1-X^2)^{(q-1)/2}}$$

$$+\frac{q^3(q+1)}{2}\sum_{\substack{d|q-1\\ d\neq(q-1)/2,q-1}}\frac{\phi((q-1)/d)}{(1-X^{(q-1)/d})^d(1-(-1)^dX)}$$

$$+\frac{q^3(q-1)}{2}\sum_{\substack{d|q+1\\ d\neq(q+1)/2,q+1}}\frac{\phi((q+1)/d)(1-(-1)^dX)}{(1-X^{(q+1)/d})^d}$$

$$+\frac{q(q^2-1)}{2}\left(\frac{f_+(X)+f_-(X)}{(1-X^p)^{2q/p}}\right)+q^2\frac{(q^2-1)}{2}\left(\frac{(g_+(X)+g_-(X))(1-X^2)}{(1-\varepsilon X)^2(1-X^{2p})^{q/p}}\right),$$

and $S(X)$ is given by

$$S(X)=p(q^2-1)\frac{(q^2-q+1)}{(1-X^p)^{q/p}}\quad \text{if } p>3,$$

while for $p=3$, $S(X)$ is given by equation (*).

Notice that $\Phi(X)$ represents the contribution to the Molien series of $\tilde{G}$ from a set of coset representatives for $\mathbb{Z}$ in $\tilde{G}$, except for those classes appearing in the last three lines of the table.

Recall that $G=\langle c_q\rangle\times\tilde{G}$.

THEOREM 6.11. *The Molien series for G with respect to ρ is given by*

$$\Phi_G(X)=\tfrac{1}{4}(\Phi_{\tilde{G}}(X)+\Phi_{\tilde{G}}(-X)+\Phi_{\tilde{G}}(iX)+\Phi_{\tilde{G}}(-iX))$$

if $q\equiv 3 \bmod 4$, and

$$\Phi_{\tilde{G}}(X)=\tfrac{1}{2}(\Phi_{\tilde{G}}(X)+\Phi_{\tilde{G}}(-X))$$

if $q\equiv 1 \bmod 4$.

The proof of this result follows from Theorem 6.10 in the same way that the proof of Theorem 6.9 follows from Theorem 6.8 and will be omitted.

7. The Molien series (characteristic two). From § 5, the group G has the form $G=O^+(V)E$. Let ρ denote either of the two extensions of ρ_1 to G guaranteed by Theorem 5.2. Let $x=\begin{pmatrix}0&1\\1&0\end{pmatrix}$ so that $O^+(V)=\langle x\rangle H$ and $G=\langle x\rangle HE$. In calculating the Molien series of G with respect to ρ, it is convenient to partition G into E, $HE-E$ and $G-HE=xHE$, and to calculate the contribution to the series from each of these subsets.

LEMMA 7.1. *The elements of E have order at most 4. The number of elements of each type and their contributions to the Molien series is given in Table 3.*

TABLE 3

Type of element g	Number of such elements	$\det(I-X\rho(g))$
identity element: (0, 0, 0)	1	$(1-X)^q$
central involution: (0, 0, 1)	1	$(1+X)^q$
noncentral involution	q^2+q-2	$(1-X^2)^{q/2}$
element of order 4	q^2-q	$(1+X^2)^{q/2}$

Proof. The first two lines of the table are clear. Let $e=(v_1, v_2, a)\in E-\mathbb{Z}(E)$. Then $e^2=(0, 0, Q(v_1, v_2))=(0, 0, \operatorname{tr}(v_1v_2))$.

Suppose first e is a noncentral involution. If $v_1=0$ then there is no restriction on $v_2\in F$ or $a\in GF(2)$, other than that $v_2\neq 0$. If $v_1\neq 0$ then $a\in GF(2)$ is arbitrary but $v_2\in V_1^{-1}$ ker (tr). Therefore, there are $(q-1)\cdot 2+(q-1)\cdot 2\cdot (q/2)=q^2+q-2$, noncentral involutions in E.

Let ζ be the character afforded by ρ_E. As $\zeta(1)^2=|E:\mathbb{Z}(E)|$, ζ is fully ramified over $\mathbb{Z}(E)$ and hence vanishes on $E-\mathbb{Z}(E)$. In particular, $\rho(e)$ has $q/2$ eigenvalues equal to 1 and $q/2$ eigenvalues equal to -1, and the third line of the table follows.

The remaining q^2-q elements of $E-\mathbb{Z}(E)$ have order 4. If e is one of these elements, then $\rho(e^2)=\rho((0, 0, 1))=-I$ and hence the eigenvalues of $\rho(e)$ are i and $-i$. As already noted, $\zeta(e)=0$ and this means that i and $-i$ appear with equal multiplicity $q/2$. The last line of the table is now verified, and the proof of the lemma is complete. □

Notice that when $q=2$, H is the identity group and $HE-E$ is empty. There is no contribution to the Molien series in this case.

LEMMA 7.2. *Assume $q>2$ and $h\in H$, $h\neq 1$. If the order of h is d then the coset Eh contains q^2 elements of order d and q^2 elements of order $2d$. If $y\in Eh$ has order d then* $\det(I-X\rho(y))=(1-X^d)^{(q-1)/d}(1-X)$, *while if y has order $2d$ this polynomial is* $(1+X^d)^{(q-1)/d}(1+X)$.

Proof. Let $\mathbb{Z}(E)=\langle z\rangle$. The group H acts Frobeniously (fixed point freely) on $E/\langle z\rangle$, so any element of Eh has order d mod $\langle z\rangle$. The map $y\mapsto yz$ exchanges elements of order d in Eh with those of order $2d$. Moreover, if y has order d, then y is conjugate to h. Since $\rho|_{\langle h\rangle}$ is similar to $1_{\langle h\rangle}+((q-1)/d)$ (regular representation of $\langle h\rangle$), we have $\det(I-X\rho(y))=(1-X^d)^{(q-1)/d}(1-X)$. If y has order $2d$ then yz has order d and is conjugate to h. Hence, $\rho(y)$ is similar to $\rho(hz)=-\rho(h)$, and $\det(I-X\rho(y))=\det(I+X\rho(h))=(1+X^d)^{(q-1)/d}(1+X)$ follows. □

For the remainder of this section x will always denote $\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$. Before analyzing the elements of xHE, it is convenient to record some useful relations which hold in G.

LEMMA 7.3. *Let $x=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$, $h, h_1\in H$, $t, t_1\in T$, $v, v_1\in F$ and $a\in GF(2)$. Then:*

(1) $(xh_1E)^h=xh^2h_1E$.

(2) $(xt_1A)^t=xtt_1A$.

(3) $(x(0, v, a))^2=(v, v, \operatorname{tr}(v^2))=(v, v, \operatorname{tr}(v))$.

(4) $(v, v, a)^2=(0, 0, \operatorname{tr}(v^2))=(0, 0, \operatorname{tr}(v))$.

(5) $(v, 0, 0)^{-1}(x(0, v, a))(v, 0, 0) = (x(0, v, a))^{-1}$.

(6) $(v_1, v_1, 0)^{-1}(x(0, v, a))(v_1, v_1, 0) = x(0, v\ a + \mathrm{tr}\,(v_1(v+1)))$.

The proof of this lemma follows easily from the definition of the multiplication of elements of G as defined in § 5, and will be omitted. Notice that in (3) and (4) we used $\mathrm{tr}\,(v^2) = \mathrm{tr}\,(v)$, as v^2 is an algebraic conjugate of v.

LEMMA 7.4. *The elements of xHE have order* 2, 4 *or* 8.

Tables 4 and 5 determine the number of elements of xHE of each type and their contributions to the Molien series.

TABLE 4
(q *a square*)

Order of element	Number of such elements	$\det(I - X\rho(g))$
2	$q(q-1)\cdot 2$	$(1-X^2)^{q/2}$
4, type 1	$q(q-1)\cdot(q-4)$	$(1-X^4)^{q/4}$
4, type 2	$q(q-1)$	$(1+X^2)^{q/4}(1-X)^{\alpha}(1+X)^{\beta}$
4, type 3	$q(q-1)$	$(1+X^2)^{q/4}(1-X)^{\beta}(1+X)^{\alpha}$
8	$q(q-1)\cdot q$	$(1+X^4)^{q/4}$
		$\alpha, \beta = q/4 \pm \sqrt{q}/2$

TABLE 5
(q *not a square*)

Order of element	Number of such elements	$\det(I - X\rho(g))$
2	$q(q-1)\cdot 2$	$(1-X^2)^{q/2}$
4	$q(q-1)\cdot(q-2)$	$(1-X^4)^{q/4}$
8, type 1	$q(q-1)\cdot(q-2)$	$(1+X^4)^{q/4}$
8, type 2	$q(q-1)$	$(1-\sqrt{2}X+X^2)^{\alpha}(1+\sqrt{2}X+X^2)^{\beta}$
8, type 3	$q(q-1)$	$(1-\sqrt{2}X+X^2)^{\beta}(1+\sqrt{2}X+X^2)^{\alpha}$
		$\alpha, \beta = q/4 \pm \sqrt{q}/8$

Proof. The group H acts by conjugation on the set $xHE = \bigcup_{h\in H} xhE$ and, by Lemma 7.3 (1), H is transitive and regular on the sets appearing in the union. Moreover, T acts by conjugation on the set $xE = \bigcup_{t\in T} xtA$ and Lemma 7.3 (2) shows this action also, to be transitive and regular. Since conjugation preserves orders of elements, only elements of xA need be considered.

Let $y = x(0, v, a) \in xA$. By Lemma 7.3 (3), y is an involution if and only if $v = 0$. Hence, x and xz are the only involutions of xA. Moreover, $\rho(x)$ and $\rho(xz)$ are M_1 and $-M_1$ in some order, by Theorem 5.2. Now, $\mathrm{trace}\,(M_1) = q^{-1/2}\sum_{r\in F}\lambda(r^2) = q^{-1/2}\sum_{r\in F}\lambda(r) = 0$, so each of these two matrices have 1 and -1 as eigenvalues with multiplicity $q/2$. The first line of both tables now follows.

Now suppose $y = x(0, v, a)$ has order 4. By Lemma 7.4 (3) and (4) we have $v \neq 0$ and $v \in \ker(\mathrm{tr})$. Hence, there are $(q-2)$ elements of order 4 in xA. These elements are conjugate to their inverses (by equation (5) of the same lemma) and when $v \neq 1$, Lemma 7.3 (6) shows that y is conjugate to yz, where $z = (0, 0, 1)$. Thus, when $v \neq 1$, the eigenvalues 1, -1, i and $-i$ appear with equal multiplicity in $\rho(y)$. In this case

$\det(I - X\rho(y)) = (1-X^4)^{q/4}$. When q is not a square, $1 \notin \ker(\mathrm{tr})$ so $v \neq 1$ is automatically satisfied, and the second line of Table 5 follows.

Assume that q is a square. Then there are $q-4$ elements of order 4 in xA which satisfy $\det(I - X\rho(y)) = (1-X^4)^{q/4}$. These correspond to the elements of order 4 in xHE of "type 1" in Table 4. The remaining elements of order 4 in xA are $y = x(0, 1, 0)$ and $yz = x(0, 1, 1)$, and these will correspond to the elements of order 4 in xHE of types 2 and 3. Assume that ρ is chosen so that $\rho(x) = M_1$, that is, $c = 1$ in the situation covered by Theorem 5.2. From $\rho|_E = \rho_1|_E$ and Theorem 5.1, we have $\rho(y) = \rho(x)\rho(0, 1, 0) = M_1 E_1$. Hence, $\mathrm{trace}(\rho(y)) = \sum_{r,s \in F} q^{-1/2}\lambda(rs)\delta_{s,r}\lambda(r) = \sum_{r \in F} q^{-1/2}\lambda(r^2)\lambda(r) = \sum_{r \in F} q^{-1/2} = q^{1/2}$. Let α, β and γ be the multiplicities of 1, -1 and i as eigenvalues of $\rho(y)$. Then $\alpha + \beta + 2\gamma = q$ and $\alpha - \beta = q^{1/2}$.

Now $y^2 = (1, 1, \mathrm{tr}(1)) = (1, 1, 0)$. By Theorem 5.1 the (r, r) entry of $\rho(y^2)$ is $\delta_{r+1,r}\lambda_1(r)\mu_0(0) = 0$, so $\mathrm{trace}(\rho(y^2)) = 0$. Hence, $\alpha + \beta - 2\gamma = 0$. Solving for α, β and γ, we have

$$\alpha = \frac{q}{4} + \frac{\sqrt{q}}{2}, \qquad \beta = \frac{q}{4} - \frac{\sqrt{q}}{2}, \qquad \gamma = \frac{q}{4}.$$

These equations imply $\det(I - X\rho(y)) = (1+X^2)^{q/4}(1-X)^{\alpha}(1+X)^{\beta}$ and $\det(I - X\rho(yz)) = \det(I + X\rho(y)) = (1+X^2)^{q/4}(1-X)^{\beta}(1+X)^{\alpha}$ and the lines corresponding to elements of order 4 in Table 4 are now verified.

Finally, suppose $y = x(0, v, a)$ has order 8. By Lemma 7.3 (3) and (4), we have $v \notin \ker(\mathrm{tr})$, and that there are q elements of order 8 in xA. As in the case of order 4, y is conjugate to y^{-1}, and when $v \neq 1$, y is conjugate to yz.

Assume first that q is a square. Then $1 \in \ker(\mathrm{tr})$ and $v \neq 1$ is automatically satisfied. Since $\rho(y^4) = -I$, the eigenvalues of $\rho(y)$ are primitive eighth roots of unity, and, from the previous paragraph, they appear with equal multiplicity $q/4$. The last line of Table 4 is now immediate.

Assume then q is not a square. Then there are $q-2$ elements of order 8 in xA which satisfy $\det(I - X\rho(y)) = (1+X^4)^{q/4}$ and these correspond to the elements of order 8 in xHE of type 1 in Table 5.

Consider now the elements $y = x(0, 1, 0)$ and $yz = x(0, 1, 1)$. These will correspond to the elements of order 8 in xHE of types 2 and 3. If $\varepsilon = \exp(\pi i/4)$ then ε is a primitive eighth root of one, and the eigenvalues of $\rho(y)$ are ε and ε^{-1} (each with multiplicity α, for example) and ε^3 and ε^{-3} (each with multiplicity β, say). As already calculated, $\mathrm{trace}(\rho(y)) = q^{1/2}$ and this leads to the system

$$2\alpha + 2\beta = q, \qquad \alpha(\varepsilon + \varepsilon^{-1}) + \beta(\varepsilon^3 + \varepsilon^{-3}) = q^{1/2}.$$

Now $\varepsilon + \varepsilon^{-1} = \sqrt{2}$ and $\varepsilon^3 + \varepsilon^{-3} = -\sqrt{2}$, so the solution is $\alpha = q/4 + \sqrt{q/8}$ and $\beta = q/4 - \sqrt{q/8}$. Hence, $\det(I - X\rho(y)) = (1-\varepsilon X)^{\alpha}(1-\varepsilon^{-1}X)^{\alpha}(1-\varepsilon^3 X)^{\beta} \cdot (1-\varepsilon^{-3}X)^{\beta} = (1-\sqrt{2}X+X^2)^{\alpha}(1+\sqrt{2}X+X^2)^{\beta}$. The last two lines of Table 5 now follow, and the proof of Lemma 7.4 is complete. □

Combining all of the above lemmas, the Molien series for G may now be written down.

THEOREM 7.5. *The Molien series for G is given as follows*:

$$\begin{aligned} 4q^2(q-1)\Phi_G(X) &= (1-X)^{-q} + (1+X)^{-q} + (q^2+q-2)(1-X^2)^{-q/2} \\ &\quad + (q^2-q)(1+X^2)^{-q/2} \\ &\quad + q^2 \sum_{1 \neq d \mid q-1} \phi(d)\{(1-X^d)^{-(q-1)/d}(1-X)^{-1} \\ &\qquad + (1+X^d)^{-(q-1)/d}(1+X)^{-1}\} + A_q(X). \end{aligned}$$

Here $A_q(X)$ is a rational function determined as follows. If q is a square, then

$$A_q(X) = 2q(q-1)(1-X^2)^{-q/2} + q(q-1)(q-4)(1-X^4)^{-q/4}$$
$$+ q(q-1)(1+X^2)^{-q/4}\{(1-X)^{-\alpha}(1+X)^{-\beta} + (1-X)^{-\beta}(1+X)^{-\alpha}\}$$
$$+ q^2(q-1)(1+X^4)^{-q/4},$$

where $\alpha, \beta = q/4 \pm \sqrt{q}/2$. If q is not a square, then

$$A_q(X) = 2q(q-1)(1-X^2)^{-q/2} + q(q-1)(q-2)(1-X^4)^{-q/4}$$
$$+ q(q-1)(q-2)(1+X^4)^{-q/4}$$
$$+ q(q-1)\{(1-\sqrt{2}X+X^2)^{-\alpha}(1+\sqrt{2}X+X^2)^{-\beta}$$
$$+ (1-\sqrt{2}X+X^2)^{-\beta}(1+\sqrt{2}X+X^2)^{-\alpha}\},$$

where $\alpha, \beta = q/4 \pm \sqrt{q/8}$.

8. Finite extensions of the linear group $\rho(\tilde{G})$. In this section G, ρ and q have the same meaning as in § 5. In particular, $\rho(G)$ is a linear group of degree q which leaves invariant the complete weight enumerator of a normalized self-dual code over $G = GF(q)$. As in § 5, q is a power of some odd prime p.

For $q > 3$, the unimodular subgroup of $\rho(G)$ is $\rho(\tilde{G})$. The goal of this section is to establish that the finite unimodular linear groups containing $\rho(\tilde{G})$ are all contained in a unique largest such group (Theorem 8.11 below). If $q = 3$, the commutator subgroup of $\rho(\tilde{G})$ is the unimodular subgroup of $\rho(G)$, however a similar result holds for this group (Theorem 8.12 below). The proof of this result is omitted here since it follows from the classification of linear groups of degree three which was already known to Blichfeldt [3].

Recall that $\tilde{G}$ has the form SE, where E is a normal extraspecial p-group and $S = SL(2, F)$ acts on E as a group of automorphisms, centralizing $\mathbb{Z}(E)$. The full automorphism group of E which centralizes $\mathbb{Z}(E)$ is isomorphic to a split extension of the symplectic group $\mathrm{Sp}(2n, p)$ by the inner automorphisms of E. Here $q = p^n$. The semi-direct product $G_1 = \mathrm{Sp}(2n, p) \ltimes E$ may be formed, and it is not hard to establish that the representation ρ extends to G_1. A proof of this fact is contained in the proof of [12, Thm. 4.7]. For $q > 3$, G_1 is perfect and hence $\rho(G_1)$ is a finite unimodular group containing $\rho(\tilde{G})$.

A further extension of G_1 is necessary. Let $\mathbb{Z}_q$ denote a cyclic group of order q and set $G_2 = G_1 Y \mathbb{Z}_q$. This is a central product in which $\mathbb{Z}(G_1)$ is identified with the unique subgroup of order p in $\mathbb{Z}_q$. Clearly, ρ extends to G_2 in such a way that $\mathbb{Z}_q$ is represented by scalar matrices. Notice that if $q = p$ then $G_2 = G_1$.

The linear group $\rho(G_2)$ will turn out to be the unique maximal finite subgroup of $SL(q, \mathbb{C})$ containing $\rho(\tilde{G})$. For convenience, G_2 and its various subgroups will be identified with subgroups of $SL(q, \mathbb{C})$ via the representation ρ. This notation will be fixed throughout the section. We also assume for the remainder of this section that $q > 3$.

LEMMA 8.1. *The group $\tilde{G} = SE$ is a primitive linear group.*

Proof. By Theorem 5.1, E (and hence SE) is an irreducible linear group. Suppose that it is not primitive, and that the representation is induced from the proper subgroup R. Clearly $R \supseteq \mathbb{Z}(\tilde{G})$. Since E is irreducible, $\tilde{G} = ER$ and $R_0 = R \cap E$ is a proper subgroup of E containing $\mathbb{Z}(E)$. Hence, $R_0 \trianglelefteq E$ and clearly $R_0 \trianglelefteq R$, so $R_0 \trianglelefteq \tilde{G}$. Since S acts irreducibly on $E/\mathbb{Z}(E)$, R_0 must be either E or $\mathbb{Z}(E)$. However, R_0 is a proper subgroup of E with index $\leqq q$, and this contradiction proves the lemma. □

LEMMA 8.2. *If $\mathbb{Z}$ denotes the group of scalar matrices in $GL(q, \mathbb{C})$ then $\mathbb{C}_{GL(q,\mathbb{C})}(E/\mathbb{Z}(E)) = E\mathbb{Z}$. In particular $\mathbb{C}_{SL(q,\mathbb{C})}(E/\mathbb{Z}(E)) = E\mathbb{Z}_q$.*

Proof. Clearly, $E\mathbb{Z} \leqq \mathbb{C}_{GL(q,\mathbb{C})}(E/\mathbb{Z}(E))$. Let $g \in \mathbb{C}_{GL(q,\mathbb{C})}(E/\mathbb{Z}(E))$. If $x_1, \cdots, x_k$ is a basis for $E/\mathbb{Z}(E)$ then $g^{-1}x_i g = \zeta_i x_i$ for each i, where $\zeta_i \in \mathbb{Z}(E)$. For a given g then, one of at most ρ^k sequences $(\zeta_1, \zeta_2, \cdots, \zeta_k)$ is determined. However, since E acts on itself as a group of inner automorphisms, and $|E: \mathbb{Z}(E)| = p^k$, every such sequence actually appears. Hence, there exists $e \in E$ such that $g^{-1}x_i g = e^{-1}x_i e$ for every i. Thus, ge^{-1} centralizes E and since E is irreducible, $ge^{-1} = \zeta$ for some scalar matrix ζ. Hence $g = e\zeta \in E\mathbb{Z}$. If in addition $\det g = 1$, then, since $\det e = 1$, we must have $\zeta \in \mathbb{Z}_q$ and the lemma follows. □

LEMMA 8.3. $\mathbb{N}_{SL(q,\mathbb{C})}(E) = G_2$.

Proof. Clearly, G_2 normalizes E. Suppose $g \in \mathbb{N}_{SL(q,\mathbb{C})}(E)$. Then conjugation by g is an automorphism of E centralizing $\mathbb{Z}(E)$. Now $\mathbb{C}_{\mathrm{Aut}(E)}(\mathbb{Z}(E)) \simeq \mathrm{Sp}(2n, p)E/\mathbb{Z}(E)$ so there exists $h \in \mathrm{Sp}(2n, p)E$ such that gh^{-1} centralizes E. Since $\det h = 1$, we have $gh^{-1} \in \mathbb{C}_{SL(q,\mathbb{C})}(E) = \mathbb{Z}_q$ and so $g \in G_1\mathbb{Z}_q = G_2$, as desired. □

LEMMA 8.4. *Let X be a finite subgroup of $SL(q, \mathbb{C})$ containing $\tilde{G}$ and assume that $E \leqq O_p(X)$. Then $O_p(X) \leqq E\mathbb{Z}_q$ and $E \trianglelefteq X$. In particular, $X \leqq G_2$.*

Proof. Neither the hypotheses, nor the conclusion is affected if X is replaced by $X\mathbb{Z}_q$. Hence, we may assume that $\mathbb{Z}_q \leqq X$, and so $\mathbb{Z}_q \leqq O_p(X)$. Thus $E\mathbb{Z}_q \leqq O_p(X)$. If $E\mathbb{Z}_q < O_p(X)$ then $E\mathbb{Z}_q < \mathbb{N}_{O_p(X)}(E\mathbb{Z}_q)$. Now S normalizes these last two groups. Choose Y so that $E\mathbb{Z}_q < Y \leqq \mathbb{N}_{O_p(X)}(E\mathbb{Z}_q)$, and $Y/E\mathbb{Z}_q$ is an S-composition factor. Therefore, both $Y/E\mathbb{Z}_q$ and $E\mathbb{Z}_q/\mathbb{Z}_q$ are irreducible under the action of S. As Y is a p-group, $[Y, E\mathbb{Z}_q/\mathbb{Z}_q] < E\mathbb{Z}_q/\mathbb{Z}_q$ and is S-invariant. Hence, $[Y, E\mathbb{Z}_q/\mathbb{Z}_q]$ is trivial, and $[Y, E\mathbb{Z}_q] \leqq \mathbb{Z}_q$. In particular, $[Y, E] \leqq [Y, E\mathbb{Z}_q] \cap E \leqq \mathbb{Z}_q \cap E = \mathbb{Z}(E)$, and so $Y \leqq \mathbb{C}_{SL(q,\mathbb{C})}(E/\mathbb{Z}(E)) = E\mathbb{Z}_q$, where the last equality follows from Lemma 8.2. But this contradicts $E\mathbb{Z}_q < Y$ and hence $O_p(X) = E\mathbb{Z}_q$. Finally, $E = \Omega_1(E\mathbb{Z}_q) = \Omega_1(O_p(X)) \trianglelefteq X$, and $X \leqq G_2$ follows from Lemma 8.3. □

THEOREM 8.5. *If X is a finite subgroup of $SL(q, \mathbb{C})$ containing $\tilde{G}$ and if the Fitting subgroup $\mathbb{F}(X)$ is not contained in $\mathbb{Z}(X)$ then $X \leqq G_2$.*

Proof. Let r be a prime divisor of the order of $\mathbb{F}(X)$. Now X is primitive (as $\tilde{G}$ is, by Lemma 8.1), so $\mathbb{Z}(O_r(X))$ is represented by scalar matrices. But X is unimodular, so $\mathbb{Z}(O_r(X)) \leqq \mathbb{Z}_q$, and hence $O_r(X) = 1$ unless $r = p$. Thus, $\mathbb{F}(X) = O_p(X)$. The group $\tilde{G} = SE$ normalizes $O_p(X)$, so $SEO_p(X)$ is a group. Hence, $EO_p(X) = O_p(SEO_p(X)) \leqq E\mathbb{Z}_q$ by Lemma 8.4. Since S is irreducible and nontrivial on $E\mathbb{Z}_q/\mathbb{Z}_q$, all X-composition factors of $E\mathbb{Z}_q$ but one are trivial. By hypothesis, $O_p(X) = \mathbb{F}(X) \not\leqq \mathbb{Z}(X)$ and so we have $O_p(X) \cdot \mathbb{Z}_q = E\mathbb{Z}_q$. Therefore, $E = [E, S] \leqq [E\mathbb{Z}_q, S] = [O_p(X)\mathbb{Z}_q, S] = [O_p(X), S] \leqq O_p(X)$. By Lemma 8.4 again, $X \leqq G_2$. □

Notice that the main theorem of this section follows from the previous result, once it can be established that the Fitting subgroup of any finite unimodular linear group containing $\tilde{G}$ is not central. Toward this end, it is natural to consider the following situation.

Hypothesis 8.6. X is a finite subgroup of $SL(q, \mathbb{C})$ containing $\tilde{G}$ and $\mathbb{F}(X) \leqq \mathbb{Z}(X)$.

LEMMA 8.7. *Assume Hypothesis* 8.6 *holds for X. Then the generalized Fitting subgroup $\mathbb{F}^*(X)$ has the form $\mathbb{Z}(X)Y$, where Y is a quasi-simple group and $E \leqq Y$.*

Proof. By definition, $\mathbb{F}^*(X)$ is $\mathbb{F}(X)\mathbb{E}(X)$, where $\mathbb{E}(X)$ is the join of all the quasi-simple subnormal subgroups of X, say $\mathbb{E}(X) = Y_1 Y_2 \cdots Y_t$. By hypothesis, $\mathbb{F}(X) \leqq \mathbb{Z}(X)$, and so it remains to prove that $t = 1$ and $E \leqq Y_1$.

Assume that at least one of the quasi-simple subnormal subgroups of X, say Y_1, is not normalized by S. Assume the notation is chosen so that $\{Y_1, \cdots, Y_{t'}\}$ is the orbit of S containing Y_1. An irreducible constituent of $Y_1 Y_2 \cdots Y_{t'}$ is a tensor product of t' representations, each of degree at least two, and hence, $2^{t'} \leqq q$. We assume that $q > 3$

is odd, so by a theorem of Galois (see [10, Satz 8.28, p. 214]) either $t' \geqq q$ or $t' = 6$ and $q = 9$, and this contradicts $2^{t'} \leqq q$.

Hence, S normalizes each of the groups $Y_1, \cdots, Y_t$. Therefore, S is in the kernel of the action of X on $\{Y_1, Y_2, \cdots, Y_t\}$ and so $E = [E, S]$ is also in this kernel. Hence, $\tilde{G} = SE$ normalizes each Y_i.

Suppose there exists i such that $E \not\leqq Y_i$. Let $\bar{E} = E\mathbb{Z}_q/\mathbb{Z}_q$ and $\bar{Y}_i = Y_i\mathbb{Z}_q/\mathbb{Z}_q$. Define the group K_i by $K_i/\mathbb{Z}_q = \bar{K}_i = \mathbb{N}_{\bar{Y}_i}(\bar{E})$. Then $[\bar{K}_i, \bar{E}] \leqq \bar{Y}_i \cap \bar{E} = (Y_i\mathbb{Z}_q \cap E\mathbb{Z}_q)/\mathbb{Z}_q$. If $Y_i\mathbb{Z}_q \cap E\mathbb{Z}_q > \mathbb{Z}_q$, then this group must equal $E\mathbb{Z}_q$, as S is irreducible on $E\mathbb{Z}_q/\mathbb{Z}_q$. But then $E\mathbb{Z}_q \leqq Y_i\mathbb{Z}_q$, and so $E = [E\mathbb{Z}_q, S] \leqq [Y_i\mathbb{Z}_q, S] = [Y_i, S] \leqq Y_i$, a contradiction. Hence, $\bar{Y}_i \cap \bar{E} = \bar{1}$ and $\bar{K}_i$ centralizes $E\mathbb{Z}_q/\mathbb{Z}_q$. Now $[K_i, E] \leqq \mathbb{Z}_q \cap E = \mathbb{Z}(E)$, so that $K_i \leqq \mathbb{C}_{SL(q,\mathbb{C})}(E/\mathbb{Z}(E))$. By Lemma 8.3, then, $K_i \leqq E\mathbb{Z}_q$. However, $K_i \leqq Y_i$ so $K_i \leqq E\mathbb{Z}_q \cap Y_i \leqq E\mathbb{Z}_q \cap Y_i\mathbb{Z}_q = \mathbb{Z}_q$. Hence, $\bar{K}_i = \bar{1}$, proving that $\mathbb{N}_{\bar{Y}_i}(\bar{E}) = \bar{1}$.

In particular, $p \nmid |\bar{Y}_i|$ and hence $\bar{Y}_i = Y_i\mathbb{Z}_q/\mathbb{Z}_q = Y_i \times \mathbb{Z}_q/\mathbb{Z}_q \simeq Y_i$. Now SEY_i is a primitive group of degree q and $Y_i \trianglelefteq SEY_i$, so Y_i has an irreducible constituent whose degree is divisible by p. This contradicts $p \nmid |Y_i|$ and shows $E \leqq Y_i$ for every i.

If $t \geqq 2$ then $1 \neq \mathbb{Z}(E) = [E, E] \leqq [Y_1, Y_2] = 1$, a contradiction. Thus, $t = 1$ and $E \leqq Y_1$, completing the proof of Lemma 8.7. □

The following is an improvement of Lemma 8.7.

LEMMA 8.8. *Assume Hypothesis* 8.6 *holds for* X. *Then* $\mathbb{F}^*(X) = \mathbb{Z}(X)Y$, *where* Y *is quasi-simple. If* $\bar{Y} = Y/\mathbb{Z}(Y)$ *and* $\mathrm{Aut}(\bar{Y})/\mathrm{Inn}(\bar{Y})$ *is solvable, then* $\tilde{G} = SE \leqq Y$.

Proof. By Lemma 8.7, it suffices to prove that $S \leqq Y$. Now S acts on Y and $\bar{Y} = Y/\mathbb{Z}(Y)$, as a group of automorphisms. Since S is perfect and $\mathrm{Aut}(\bar{Y})/\mathrm{Inn}(\bar{Y})$ is solvable, S acts as a group of inner automorphisms of $\bar{Y}$. Let $s \in S$. Then there exists $y \in Y$ such that $s^{-1}xs \equiv y^{-1}xy \bmod \mathbb{Z}(Y)$ for all $x \in Y$. Hence, $[sy^{-1}, Y] \leqq \mathbb{Z}(Y)$, and so $[sy^{-1}, Y, Y] = 1$. By the three subgroups lemma [10, Lemma , p. 257], $[Y', sy^{-1}] = 1$. But $Y' = Y$, so $sy^{-1} \in \mathbb{C}_{SL(q,\mathbb{C})}(Y) \leqq \mathbb{C}_{SL(q,\mathbb{C})}(E) = \mathbb{Z}_q$. Hence, $s \in y\mathbb{Z}_q \subseteq Y\mathbb{Z}_q$, so $S \leqq Y\mathbb{Z}_q$. Therefore, $S = S' \leqq (Y\mathbb{Z}_q)' = Y' \leqq Y$, as desired. □

To complete the proof of the main theorem of this section, it suffices to show $\mathbb{F}^*(X)$ does not have the form given in Lemma 8.7 for every choice of the simple group $Y/\mathbb{Z}(Y)$. We shall assume the classification of finite simple groups is complete and consists of the alternating groups, the Chevalley groups and their twisted types, and the 26 exceptional groups, as listed in [1] for example. Since the Schreier conjecture holds for these groups, Lemma 8.8 is applicable. Moreover, notice that the conclusion of Lemma 8.7 implies that $\mathbb{Z}(E) \leqq \mathbb{Z}(Y)$, since $\mathbb{Z}(E)$ consists of scalars. Hence, the p-part of the Schur multiplier of the simple group $\bar{Y} = Y/\mathbb{Z}(Y)$ is nontrivial. A complete list of Schur multipliers for simple groups appears in [8].

There are four infinite families of simple groups with the Schur multiplier divisible by an odd prime p. These are

$$PSL(n, r), \quad p|(n, r-1),$$
$$PSU(n, r), \quad p|(n, r+1),$$
$$\overline{E_6(r)}, \quad p = 3|(r-1),$$
$$\overline{{}^2E_6(r)}, \quad p = 3|(r+1).$$

A bar denotes the quotient of the universal Chevalley group by its center. There are twelve other exceptional groups. In each of these cases $p = 3$.

$$A_6, \quad A_7, \quad M_{22}, \quad J_3, \quad O'S,$$
$$G_2(3), \quad PSU(4, 3), \quad \mathrm{McL}, \quad \mathrm{Suz},$$
$$SO(7, 3), \quad \mathrm{Fi}_{22} = M(22), \quad \mathrm{Fi}'_{24} = M(24)'.$$

Each of these groups will be eliminated from occurring as $Y/\mathbb{Z}(Y)$ in hypothesis 8.6. The next result dispenses with the four infinite families first.

LEMMA 8.9. *Let X satisfy hypothesis* 8.6 *and let Y be as in Lemma* 8.7. *Then $\bar{Y} = Y/\mathbb{Z}(Y)$ is not one of the groups $PSL(n, r)$, $PSU(n, r)$, $\overline{E_6(r)}$, $\overline{{}^2E_6(r)}$.*

Proof. Assume that $\bar{Y}$ is one of these groups. Since S acts transitively on the nonidentity elements of $E/\mathbb{Z}(E)$, it follows that S also acts transitively on the set of Brauer characters of $E/\mathbb{Z}(E)$ over $GF(r)$, where $p|(r-1)$. Therefore, any faithful representation of $SE/\mathbb{Z}(E)$ over $GF(r)$ must have degree at least $|E: \mathbb{Z}(E)|-1 = q^2-1$. Since $SE/\mathbb{Z}(E)$ embeds in $\bar{Y}$ by Lemma 8.8, the same result holds for $\bar{Y}$. (In fact, a similar argument shows that the degree of any nonprincipal character of $\bar{Y}$ is at least q^2-1.)

If $\bar{Y}$ is $PSL(n, r)$ or $\overline{E_6(r)}$ then $\bar{Y}$ acts as a group of automorphisms on the associated Lie algebra over $GF(r)$. In the first case, the Lie algebra consists of the n^2-1 dimensional space of $n \times n$ matrices of trace 0, and in the second case, it is the 78-dimensional Lie algebra of type E_6. Since $p|(r-1)$ we have

$$q^2-1 \leqq n^2-1, \quad \text{if } \bar{Y} \simeq PSL(n, r),$$

and

$$q^2-1 \leqq 78, \qquad \text{if } \bar{Y} \simeq \overline{E_6(r)}.$$

If $\bar{Y}$ is $PSU(n, r)$ or $\overline{{}^2E_6(r)}$ then $\bar{Y}$ is isomorphic to a subgroup of $PSL(n, r^2)$ or $\overline{E_6(r^2)}$. Moreover, in this case p divides $(r+1)$ which in turn divides r^2-1. Therefore, the inequalities remain valid for $PSU(n, r)$ and $\overline{{}^2E_6(r)}$.

If $\bar{Y}$ is $\overline{E_6(r)}$ or $\overline{{}^2E_6(r)}$ then $q^2-1 \leqq 78$ which implies $q < 9$. But for these groups we have $p = 3$, and since $q \neq 3$ is a power of p, we must have $q \geqq 9$. This contradiction shows that $\bar{Y}$ is $PSL(n, r)$ or $PSU(n, r)$ and $q \leqq n$.

Assume first $\bar{Y}$ is $PSL(n, r)$, where $p|(n, r-1)$. Let $R = \left\{\begin{pmatrix} M & 0 \\ v & 1 \end{pmatrix} | M \in SL(n-1, r), v \in GF(r)^{n-1}\right\}$ and $U = \left\{\begin{pmatrix} I & 0 \\ v & 1 \end{pmatrix} | v \in GF(r)^{n-1}\right\}$. As R does not contain any scalar matrices other than the identity, R is isomorphic to a subgroup of both Y and $\bar{Y}$. This remark is also valid when $n = 3$ and $r = 4$, where the corresponding group has an exceptional multiplier. Moreover $U \lhd R$, and since $n-1 \geqq 2$, R acts transitively on the nonidentity elements of U. Hence, R is transitive on the set of nonprincipal characters of U. In particular, any faithful character of R (and hence of Y) when restricted to U must contain all of the nonprincipal characters of U as constituents. Since Y is a linear group of degree q, this implies $r^{n-1}-1 \leqq q$.

Combining this with the previous inequality gives $r^{n-1}-1 \leqq n$. However, $p|(r-1)$ so $r \geqq 4$, and the inequality forces $n = 1$, a contradiction.

It remains to consider the case $\bar{Y} \simeq PSU(n, r)$, where $p|(n, r+1)$. If $k = [n/2]$ then the map which sends M to

$$\begin{pmatrix} M & 0 \\ 0 & (\bar{M}^T)^{-1} \end{pmatrix}$$

is an embedding of $SL(k, r^2)$ in $SU(2k, r)$, the unitary group being defined with respect to the form whose matrix is $\begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}$. As $2k \leqq n$, we have an embedding of $SL(k, r^2)$ in $SU(n, r)$.

Assume first that $n \geqq 6$ so that $k \geqq 3$. We already observed that $SL(k, r^2)$ contains a subgroup R, which does not contain any scalar matrices other than the identity, and

which has the property that any faithful character of R has degree at least $(r^2)^{k-1}-1$. Since R embeds in the linear group Y, we conclude $r^{2k-2}-1 \leqq q$. However, $q \leqq n$ and this implies $2^{n-3} \leqq r^{n-3} \leqq n+1$. This forces the contradiction $n<6$.

Assume then $n \leqq 5$. Now $q \leqq n$ and since $q \neq 3$ this leads to the single case $p=q=n=5$. We still have an embedding of $SL(2, r^2)$ in $SU(5, r)$ and the smallest degree of any nonprincipal irreducible character of $SL(2, r^2)$ is $(r^2-1)/2$, if r is odd and (r^2-1), if r is a power of 2. In any case, we have $(r^2-1)/2 \leqq q=5$, which forces $r \leqq 3$. But $p=5|(r+1)$ and this contradiction eliminates the groups $PSU(n, r)$. □

LEMMA 8.10. *Let X satisfy hypothesis* 8.6 *and let Y be as in Lemma* 8.7. *Then $\bar{Y}=Y/\mathbb{Z}(Y)$ is not one of the groups*:

$$A_6,\quad A_7,\quad M_{22},\quad J_3,\quad O'S,$$

$$G_2(3),\quad PSU(4,3),\quad \text{McL},\ \text{Suz},$$

$$SO(7,3),\quad \text{Fi}_{22}=M(22),\quad \text{Fi}'_{24}=M(24)'.$$

Proof. Assume that $\bar{Y}$ is one of these groups. Then Lemma 8.8 applies, so that $SE/\mathbb{Z}(E)$ embeds in $\bar{Y}$. Moreover, in each of these cases $p=3$ so that $q \geqq 9$. Since the highest power of 3 dividing the order of any of the five groups in the first line is at most 5, $SE/\mathbb{Z}(E)$ cannot be embedded as a subgroup of any of these by Lagrange's theorem.

The highest power of 3 dividing the order of any of the four groups in the second line is 6 or 7. Hence, if $SE/\mathbb{Z}(E)$ is embedded in any of these, then $q=9$. The first group, $G_2(3)$, may be eliminated as $5 \nmid |G_2(3)|$. We have already noticed (first paragraph to the proof of Lemma 8.9) that any faithful representation of $SE/\mathbb{Z}(E)$, and hence of $\bar{Y}$, has degree at least $q^2-1=80$. Now McL appears as a subgroup of Conway's group $\cdot O$ and hence has a faithful representation of degree 24. Thus, $SE/\mathbb{Z}(E)$ cannot be a subgroup of McL. Since $PSU(4, 3) \leqq$ McL, $SE/\mathbb{Z}(E)$ cannot be embedded in $PSU(4, 3)$, either. Suppose now SE is a subgroup of $\widehat{\text{Suz}}$ (the threefold cover of Suz). Now Suz contains the chain

$$\text{Suz} \geqq G_2(4) \geqq SL(3,4) \geqq R=\left\{\begin{pmatrix} M & 0 \\ v & 1 \end{pmatrix} | M \in SL(2,4),\ v \in GF(4)^2\right\}.$$

Moreover, as a Sylow 3-subgroup of R is cyclic, the 3-part of the multiplier of R is trivial, so that R embeds as a subgroup of $\widehat{\text{Suz}}$. We already have seen that the smallest degree of a faithful representation of R is $4^2-1=15$. Thus, $\widehat{\text{Suz}}$ cannot be a linear group of degree $q=9$.

The highest power of 3 dividing the orders of $SO(7, 3)$ and Fi_{22} is 3^9, so if $SE/\mathbb{Z}(E)$ is embedded in either of these, then $q=9$ or 27.

Suppose first that Y is the threefold cover of $SO(7, 3)$, say $\widehat{SO(7,3)}$. Now $SO(7,3) \geqq SO^+(6,3) \simeq SL(4,3)$, and, since 3 does not divide the order of the Schur multiplier of $SL(4, 3)$, this last group appears as a subgroup of $Y=\widehat{SO(7,3)}$. As usual, let $R=\left\{\begin{pmatrix} M & 0 \\ v & 1 \end{pmatrix} | M \in SL(3,3),\ v \in GF(3)^3\right\}$. Then any faithful character of R (and hence Y) has degree $\geqq 3^3-1=26$. Thus $q=27$. Moreover, any faithful character of $SE/\mathbb{Z}(E)$ (and hence of $SO(7, 3)$) has degree $\geqq q^2-1=728$. However, the index of $SE/\mathbb{Z}(E)$ in $SO(7, 3)$ is 640, and this provides a faithful representation of degree <728, a contradiction.

Assume then that Y is the threefold cover of Fi_{22}, say $\widehat{\text{Fi}}_{22}$. Then $\widehat{\text{Fi}}_{22}$ has a faithful character of degree q, say ψ. Moreover, we know q is 9 or 27. The character $\psi\bar{\psi}$ contains the principal character as a constituent, and $(\psi\bar{\psi}-1)$ is irreducible, as its restriction to

$SE/\mathbb{Z}(E)$ is irreducible. In particular, Fi_{22} has an irreducible character of degree $q^2-1=80$ or 728. However, the entire character table of this group is known [9], and there are no characters with these degrees.

Assume finally that Y is the threefold cover of Fi'_{24}, say $\widehat{\mathrm{Fi}'_{24}}$. Then $\widehat{\mathrm{Fi}'_{24}}$ has a faithful character of degree q, and since the 3-part of the order of Fi'_{24} is 3^{16}, we know $q \leqq 3^5$. Now $\mathrm{Fi}_{23} \leqq \mathrm{Fi}'_{24}$, and 3 does not divide the order of the Schur multiplier of Fi_{23}, hence Fi_{23} as well as Fi_{22} appear as subgroups of $\widehat{\mathrm{Fi}'_{24}}$. From the character table of Fi_{22} again, 78 is the smallest degree for any faithful character of Fi_{22}. Thus $q \geqq 78$, and hence $q=81$ or 343. In either of these two cases $q+1$ does not divide $|\mathrm{Fi}'_{24}|$ so that $SE/\mathbb{Z}(E)$ cannot be embedded as a subgroup of Fi'_{24}. □

The last two lemmas establish that there is no linear group X satisfying hypothesis 8.6. When combined with Theorem 8.5 this yields the following theorem.

THEOREM 8.11. *If X is a finite subgroup of $SL(q, \mathbb{C})$ containing $\tilde{G}$ where q is an odd prime power and $q>3$, then $X \leqq G_2$.*

For convenience, we include the corresponding result when $q=3$.

THEOREM 8.12. *If X is a finite subgroup of $GL(3, \mathbb{C})$ containing $\tilde{G}$ where $q=3$ and if the scalar matrices of X are contained in $\tilde{G}$, then $X=\tilde{G}$.*

The proof of Theorem 8.11 relies on the classification of finite simple groups, a profound result whose proof currently occupies several thousand journal pages. While it is hard to imagine that such an enormous proof is absolutely free of errors, it is generally believed that the underlying argument is correct. Nevertheless, if an oversight in the proof should lead to the existence of one or several extra sporadic simple groups, then presumably these can be eliminated by ad hoc arguments (perhaps similar to those in Lemma 8.10).

As already noted in § 2, the analogue of Theorem 8.11 is false for even q, at least when q is 2 or 4. However, if $q \geqq 8$ is a power of 2, then $G_0=G$ is a primitive linear group and hence there are only finitely many subgroups of $SL(q, \mathbb{C})$ containing G. No claim is made here however that there is a unique maximal such group containing the others.

Appendix. The Molien series for G_0 and G with respect to ρ are denoted by $\Phi_{G_0}(X)$ and $\Phi_G(X)$ respectively, and are listed here for all prime powers $q \leqq 9$. For even q, the groups G_0 and G coincide.

Unnormalized codes (odd characteristic)

$$q=3 \qquad \Phi_{G_0}(X)=\frac{(1+4X^{12}+X^{24})}{(1-X^4)(1-X^{12})^2},$$

$$q=5 \qquad \Phi_{G_0}(X)=\sum_{i=0}^{13}\frac{a_i^{(5)}X^{2i}}{(1-X^4)(1-X^6)^2(1-X^{10})^2},$$

$$q=7 \qquad \Phi_{G_0}(X)=\sum_{i=0}^{21}\frac{a_i^{(7)}X^{4i}}{(1-X^4)^2(1-X^8)(1-X^{12})^2(1-X^{28})^2},$$

$$q=9 \qquad \Phi_{G_0}(X)=\sum_{i=0}^{23}\frac{a_i^{(9)}X^{2i}}{(1-X^2)(1-X^4)(1-X^8)(1-X^{10})^2(1-X^{14})^4}.$$

Normalized codes (odd characteristic)

$$q=3 \qquad \Phi_G(X)=\frac{1}{(1-X^{12})^2(1-X^{36})},$$

$$q=5 \qquad \Phi_G(X)=\sum_{i=0}^{8}\frac{b_i^{(5)}X^{10i}}{(1-X^{10})^2(1-X^{20})(1-X^{30})^2},$$

$$q=7 \qquad \Phi_G(X)=\sum_{i=0}^{11}\frac{b_i^{(7)}X^{28i}}{(1-X^{28})^3(1-X^{56})(1-X^{84})^2},$$

$$q=9 \qquad \Phi_G(X)=\sum_{i=0}^{24}\frac{b_i^{(9)}X^{6i}}{(1-X^{6})^2(1-X^{12})(1-X^{18})^3(1-X^{24})(1-X^{30})^2}.$$

Characteristic 2 ($G = G_0$)

$$q=2 \qquad \Phi_G(X)=\frac{1}{(1-X^2)(1-X^8)},$$

$$q=4 \qquad \Phi_G(X)=\frac{(1+X^{16})}{(1-X^2)(1-X^4)(1-X^6)(1-X^8)},$$

$$q=8 \qquad \Phi_G(X)=\sum_{i=0}^{20}\frac{c_i^{(8)}X^{2i}}{(1-X^2)^3(1-X^4)(1-X^8)^3(1-X^{14})}.$$

TABLE 6
Table of coefficients.

i	$a_i^{(5)}$	$a_i^{(7)}$	$a_i^{(9)}$	$b_i^{(5)}$	$b_i^{(7)}$	$b_i^{(9)}$	$c_i^{(8)}$
0	1	1	1	1	1	1	1
1	1	0	0	−1	70	−1	−2
2	0	7	0	4	1791	5	1
3	0	38	2	9	9111	27	1
4	1	109	17	9	21868	127	5
5	4	246	36	19	33015	475	0
6	10	422	89	15	33015	1345	10
7	13	636	167	2	21868	3038	6
8	10	848	278	2	9111	5819	22
9	5	1048	428		1791	9606	11
10	6	1215	590		70	13858	18
11	5	1282	704		1	17777	11
12	3	1258	760			20414	22
13	1	1095	745			20940	6
14		903	643			19322	10
15		680	504			15949	0
16		473	365			11713	5
17		289	223			7582	1
18		137	118			4303	1
19		52	56			2070	−2
20		11	23			815	1
21		2	6			264	
22			4			59	
23			1			9	
24						3	

REFERENCES

[1] M. ASCHBACHER, *The Finite Simple Groups and Their Classification*, Yale University Press, New Haven, 1980.

[2] E. R. BERLEKAMP, F. J. MACWILLIAMS AND N. J. A. SLOANE, *Gleason's theorem on self-dual codes*, IEEE Trans. Inform. Theory, IT-18 (1972), pp. 409–414.

[3] H. F. BLICHFELDT, *Finite Collineation Groups*, Univ. of Chicago Press, Chicago, IL, 1917.
[4] L. DORNHOFF, *Group Representation Theory, Part A*, Marcel Dekker, New York, 1971.
[5] P. X. GALLAGHER, *Group characters and normal Hall subgroups*, Nagoya Math. J., 21 (1962), pp. 223–230.
[6] A. M. GLEASON, *Weight polynomials of self-dual codes and the MacWilliams identities*, in 1970 Act. Congr. Int. Math. vol. 3, (1970), Gauthier-Villars, Paris, 1971, pp. 211–215.
[7] D. GORENSTEIN, *Finite Groups*, Harper and Row, New York, 1968.
[8] R. L. GRIESS, JR., *Schur multipliers of the known finite simple groups*, II, preprint.
[9] D. C. HUNT, *Character tables of certain finite simple groups*, Bull. Austral. Math. Soc., 5 (1971), pp. 1–42.
[10] B. HUPPERT, *Endliche Gruppen* I, Springer-Verlag, Berlin, 1967.
[11] I. M. ISAACS, *Character Theory of Finite Groups*, Academic Press, New York, 1976.
[12] ———, *Characters of solvable and symplectic groups*, Amer. J. Math., XCV (1973), pp. 594–635.
[13] C. JORDAN, *Mémoire sur les équations differentielles lineaires à intégrale algébrique*, J. Reine Angew. Math., 84 (1878), pp. 89–215.
[14] F. J. MACWILLIAMS, *A Theorem on the Distribution of Weights in a Systematic Code*, Bell Syst. Tech. J., 42 (1963), pp. 79–84.
[15] F. J. MACWILLIAMS, C. L. MALLOWS AND N. J. A. SLOANE, *Generalizations of Gleason's theorem on weight enumerators of self-dual codes*, IEEE Trans. Inform. Theory, IT-18 (1972), pp. 794–805.
[16] C. L. MALLOWS, V. PLESS AND N. J. A. SLOANE, *Self-dual codes over GF(3)*, SIAM J. Appl. Math., 31 (1976), pp. 649–666.
[17] T. MOLIEN, *Über die Invarianten der linear Substitutionsgruppen, Sitzungsber.*, König. Pruess. Akad. Wiss., 1897, pp. 1152–1156.
[18] V. S. PLESS, *On the uniqueness of the Golay codes*, J. Combinatorial Theory, 5 (1968), pp. 215–228.
[19] N. J. A. SLOANE, *Error-correcting codes and invariant theory: New applications of a nineteenth century technique*, Am. Math. Monthly, 84 (1977), pp. 82–107.
[20] ———, private communication.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 4, December 1981

0196-5212/81/0204-0003 $01.00/0

BROADCASTING IN TREES WITH MULTIPLE ORIGINATORS*

ARTHUR M. FARLEY† AND ANDRZEJ PROSKUROWSKI†

Abstract. Broadcasting is the information dissemination process in a communication network whereby all sites of the network become informed of a given message by calls made over lines of the network. We present an algorithm which, given a tree network and a time, determines a smallest set of subtrees covering sites of the network such that broadcast can be completed within the given time in each subtree. Information developed by the algorithm is sufficient to determine a satisfactory originator and calling scheme within each subtree.

1. Introduction. *Broadcasting* is the information dissemination process in a communication network whereby all sites of the network become informed of a given message by calls placed over lines of the network. We model a communication network by a graph $G=(V, E)$ consisting of a set V of vertices (*sites*) and a set E of edges (*lines*), each edge incident to a pair of vertices. We model processes of information dissemination by the following constraints:

(1) information is disseminated in the form of *messages*;

(2) a message is transferred by a *call* between adjacent sites;

(3) no site can participate in more than one call at any time.

The length of a message determines an associated *time unit*, being the time needed to complete a call transferring the message. As such, we will talk about the number of time units required to broadcast a message.

Broadcasting can be defined more formally as a sequence of sets $S_0 \subseteq S_1 \subseteq \cdots \subseteq S_t = V$, each set representing the sites informed of the broadcast message after time unit i, $0 \leqq i \leqq t$. For each u in $S_i - S_{i-1} (i > 0)$, there exists an adjacent site in S_{i-1}, not assigned to another site of $S_i - S_{i-1}$, which calls u during unit time i. The elements of S_0 are called the *originators* of the broadcast. The case where $|S_0| = 1$ has received considerable research attention in recent years. The minimum value of t for a given network G over all broadcasts in G is called the *broadcast time* of G; a site from which such a broadcast is possible is an element of the *broadcast center* of G. Slater, Cockayne and Hedetniemi [10] have described an algorithm for determining both parameters in an arbitrary tree network. A *tree* network is a connected, acyclic network.

Farley, Hedetniemi, Mitchell and Proskurowski [3] investigated networks having the fewest lines which allow broadcasting to be completed in the minimum possible time (i.e., $\log_2 |V|$ time units) from any site. Farley [1] discussed construction algorithms for several such minimum-time broadcast networks requiring approximately the minimum number of lines. The general problem of determining the broadcast time for a given network G has been shown to be algorithmically hard (i.e., NP-complete) by Garey and Johnson [5, p. 219]. This motivates approximate results as well as study of restricted classes of networks. Proskurowski [9] has characterized minimum broadcast trees, being rooted trees which allow broadcasting to be completed in $\log_2 |V|$ time units from the root. In [2], Farley considered broadcasting of multiple messages in completely connected networks.

In this paper, we consider a generalization of broadcasting in which several sites may originate the message (i.e., $|S_0| \geqq 1$) within a network. This could arise within practical situations in several ways. A subset of sites may be connected by a broadcast medium (i.e., radio), with message broadcast to be completed by calls over lines. The

* Received by the editors September 12, 1980, and in final form April 3, 1981. This work was supported in part by the National Science Foundation under grant ENG-79-02960.

† Department of Computer and Information Science, University of Oregon, Eugene, Oregon 97403.

situation may also arise from a hierarchical view of broadcasting within a network. A message can be seen to be broadcast through a tree of sites, each such site also being a member of a network at its "level" of the hierarchy. After broadcast in the tree is completed, informed sites originate broadcasting within these level-based networks, each such network having potentially more than one originator. The lines of the tree may be of higher speed and capacity. Networks at each level may likewise have differing communication characteristics.

We present an algorithm which, given a tree network T and a broadcast time t, determines a smallest set of subtrees covering the sites of T such that the broadcast time for each subtree is less than or equal to t. Information developed by the algorithm is sufficient to also determine a satisfactory originator and calling scheme for each subtree. A solution to our problem for $t \leqq 2$ has been given as a special case of decomposing trees into paths by Hedetniemi and Hedetniemi [6]. Our algorithm solves the problem for arbitrary $t > 0$. The algorithm is efficient, requiring time and space proportional to $|V|$, with a constant of proportionality depending linearly on t.

2. Partitioning trees by broadcast time. Trees, being acyclic connected graphs, have several properties which make them suitable for the design of efficient solution algorithms. Most important is that each vertex (and edge) separates the graph into two connected components. Therefore, there is an absence of influence between subtrees of a given vertex other than that transmitted through the vertex itself. This allows efficient algorithms to process a (current) leaf vertex, update information at its single adjacent vertex, make globally correct decisions based upon this local information and prune the leaf vertex, removing it from the tree and further consideration. We follow this paradigm in our solution algorithm for partitioning trees according to broadcast time.

The input tree is represented recursively by a *father array* [8, p. 354], which assumes an arbitrary *root* vertex. The array contains, for each nonroot vertex, a pointer to its unique father (of lesser index) on the path to the root. During execution of the algorithm, certain edges of the input tree are *cut*, disconnecting the tree and forming a subtree of the partition. At any time, the connected component of the input tree containing the root is called the *current tree*. We also refer to the *unprocessed tree*, which initially corresponds to the input tree. During each cycle of the algorithm a leaf vertex of the unprocessed tree is processed. After being processed, the leaf is pruned (i.e., removed) from the unprocessed tree, though it will remain part of the current until a cut disconnects it from the root. With each vertex v of the input tree, we associate the following information:

(i) callees(v)—a list of previously processed, adjacent neighbors, ordered according to the time that v would call them in a minimum time broadcast;
(ii) maxtime(v)—the latest time unit during which v can be called and still complete broadcasting within the required time in the subtree defined by v and subtrees of the current tree rooted by vertices on callees(v);
(iii) mintime(v)—the earliest time unit that v can be called from a (necessary) broadcast originator within a previously processed subtree of v in the current tree;
(iv) caller(v)—the adjacent, processed vertex capable of calling v with the message during time unit mintime(v) from the predetermined originator.

For each vertex v, this information is initialized as follows: callees(v) and caller(v) are set to nil (i.e., empty), maxtime(v) is set to t (as each could potentially be called during the last time unit), and mintime(v) is set to 0 (as each could potentially be an

originator). A vertex v which has a nonnil caller(v) value (has a processed subtree containing a necessary originator) is classified as *heavy*. Otherwise, v is classified as *light*. All vertices are initially light.

The processing of a leaf vertex depends on whether it is heavy or light. If a leaf vertex u is light, the attempt is made to insert u into callees(v) of its father v. If successful, maxtime(v) is updated and processing of u is complete. If u cannot be inserted, a necessary cut is introduced between u and v. If u is heavy (i.e., mintime(u) > 0), then a check is made to see whether a broadcast through u can reach all light subtrees of u (by comparing mintime(u) and maxtime(u)). If not, the subtree rooted at caller(u) is cut from the current tree; u becomes light and is processed as described above. If the light subtrees of u can be accommodated, then consideration of the father v begins. If v is light, it becomes heavy with caller(v) being u. If v is heavy, a cut is introduced, disconnecting the subtree rooted by either u or caller(v) from the current tree.

Each site u has a set of potential timeslots (1 to t) during which it can call elements of callees(u). A function *emptyslots* scans callees(u), determining the set of time slots available below a given maximum time. The maximum or minimum value returned by emptyslots is important in determining whether light subtrees can be accommodated by heavy or father sites. This outlines the approach taken by algorithm BROADCAST, which is formally defined as follows.

```
ALGORITHM BROADCAST
Input Tree T given by the array father, broadcast time t.
Output Partition of T into subtrees of broadcast time at most t represented by
       cut edges of T.
Method begin
{0. Initialize} for each vertex u do
{0.1}    begin maxtime(u) := t;
{0.2}          mintime(u) := 0;
{0.3}          callees(u) := nil;
{0.4}          caller(u) := nil end;
{1. Prune} for each leaf vertex u of unprocessed tree do
        if mintime(u) > 0 {a heavy leaf}
{1.1}      then if maxtime(u) < mintime(u) {u cannot be covered}
{1.1.1}           then begin mintime(u) := 0; cut (caller(u)); {u is light now}
                           updatefather(u)
                      end
{1.1.2}           else upminfather(u)
{1.2}      else {a light leaf} updatefather(u)
          end. {of BROADCAST}

procedure updatefather(u);
{of a light vertex}
begin if maxtime(u) = 0 {root of a broadcast}
        then {a vertex informed at time 0}
             upminfather(u)
        else upmaxfather(u)
end;

procedure upmaxfather(u);
{recompute maxtime requirements}
begin v := father(u);
      s := max[emptyslots(callees(v), maxtime(u))];
```

```
        if s = 0 then cut(u) {v cannot accommodate u}
                 else begin insert (u, callees(v), s);
                        if s ≤ maxtime(v)
                        then maxtime(v) := s − 1
                      end {v calls u at s}
end;

procedure upminfather(u);
{updates mintime info}
begin v := father(u);
      s := min[emptyslots(callees(u), t+1)];
      if mintime(v) = 0 {a light node}
          then begin caller(v) := u; mintime(v) := s+1 end {v becomes heavy}
          else if s < mintime(v) {a taller son}
                 then begin cut(caller(v));
                        caller(v) := u;
                        insert(v, callees(u), s);
                        mintime(v) := s+1
                        end
                 else cut(u)
end;

procedure cut(vertex);
{adds the edge between vertex and father (vertex) to the set of cut edges}
function emptyslots(list, min);
{returns a set of timeslots less than min at which another callee can be informed
    (inserted into list), or 0 if no such slot exists}
procedure insert(vertex, list, slot);
{inserts vertex on the list at time slot maintaining the increasing time order of list}
```

3. Correctness and complexity of the algorithm. In this section we will state and prove lemmas verifying correct computation of vertex (subtree) parameters during execution of the algorithm BROADCAST. These are used to establish correctness of the algorithm. In our arguments, we use the notions of light and heavy vertices, the current tree, and the unprocessed tree, as defined in the preceding section. Additionally, by *pruned subtree* of a vertex v we understand a subtree rooted at a pruned neighbor of v.

We first consider computation of the time parameters: maxtime and mintime.

LEMMA 1. *In the current tree S, the value* maxtime (v) *of an unprocessed vertex v equals the latest time unit (counted from the origination of the message) in which v must be informed in order to complete broadcast in its pruned light subtrees in S by time t.* Callees(v) *is the list of pruned light sons of v ordered by their maxtime values.*

Proof. During initialization, the value of maxtime(v) for each vertex v of T is set to t, which is correct for vertices having no pruned light subtree; callees(v) is initially empty. Let us assume that the values are correct just before a leaf of the unprocessed tree is pruned. If this leaf is heavy and no cut is made, then the values are left correct. A heavy leaf vertex u of the unprocessed tree whose caller must be cut off (step {1.1.1}) becomes a light vertex of the new S. Both for such a vertex and for an originally light vertex the procedure upmaxfather (called from updatefather) correctly updates the values maxtime(v) and callees(v) of the father v. These values stay unchanged if the light leaf vertex u has to be cut off. In this case, the disconnected subtree rooted at u

does not influence the maxtime or callees values of the father. Otherwise, the vertex u is inserted at an appropriate place in the list of callees of v and, if its calling time is now the earliest, it redefines the maxtime of v. □

LEMMA 2. *In a current tree S, a heavy vertex cannot have two pruned sons which are heavy or which have maxtime equal to zero.*

Proof. The broadcast time for a subtree of pruned vertices of S rooted at vertex u is greater than t if u is heavy, or equal to t if maxtime$(u) = 0$. Thus, the message to be broadcast to u cannot originate outside of its subtree as implied by a heavy brother in S. In the procedure upminfather such a situation is prevented by cutting off one of the two heavy vertices. □

LEMMA 3. *In the current tree S, the value of* mintime (v) *of a heavy vertex v equals the earliest time during which v can be informed of a message originated by a previously determined vertex in a pruned subtree of v. The heavy neighbor of v supplying this message is* caller(v).

Proof. All vertices of T are initially light. A vertex v becomes heavy when one of its pruned sons u in S has maxtime equal to zero or is heavy. In both cases, the procedure upminfather is invoked to update mintime(v) and caller(v) according to the parameters of u. Let us assume that the values are correct just before u is pruned. In upminfather, the earliest available timeslot of u is determined. If v is light, then caller(v) is correctly set to u and mintime(v) is accordingly set. If v is heavy, then this timeslot is compared to the current value of mintime(v) and a cut minimizing the resultant value of mintime(v) is made. If the current caller(v) is cut, then the values are appropriately updated according to parameters of u; otherwise they remain unchanged. □

LEMMA 4. *In the current tree S, a pruned vertex v can have a heavy son u only if u can call v at or before* maxtime(v)

Proof. By Lemma 1, we know it requires $t - \text{maxtime}(v)$ time units to complete broadcasting from v to its light, pruned subtrees in S. Therefore, a broadcast cannot be completed by time t if v is not informed prior to or at maxtime(v). When heavy son u cannot inform v prior to or at maxtime(v), the necessary cut is made in step {1.1.1}. □

THEOREM 1. *Algorithm* BROADCAST *computes a minimum-size partition of tree T such that a message can be broadcast from a single originator in each subtree within time t.*

Proof. Let $b(T, t)$ be this minimum size and $c(T, S)$ be the number of invocations of procedure *cut* in the algorithm on T, when the current tree is S. We have to prove that $c(T, \Phi) = b(T, t) - 1$. This will be shown by establishing the invariant value of $c(T, S) + b(S, t)$ during the execution. Indeed, the current tree changes only when *cut* is invoked in one of three cases: (i) when pruning a heavy vertex which cannot be accommodated (called) from its heavy descendant in S, (ii) when encountering two heavy sons in upminfather, or (iii) when a light vertex cannot be informed by its father in the required time (in upmaxfather). In all these cases, a cut has to be made according to Lemmas 2, 4 and 1, respectively. The cut results in a heavy vertex with smallest possible value of mintime ((ii)), or in a light vertex with largest possible value of maxtime ((i) or (iii)). This ensures that the new current tree, S', has the minimum value of $b(S'', t)$ over all subtrees S'' of S such that the cutoff subtree $S - S''$ has a broadcast time at most t. Thus, $b(S', t) = b(S, t) - 1$ and $c(T, S') = 1 + c(T, S)$. Therefore, $c(T, S') + b(S', t) = 1 + c(T, S) + b(S, t) - 1 = c(T, S) + b(S, t)$. Initially, $S = T$, and this constant value $c(T, T) + b(T, t) = b(T, t)$. After the final cut has been made, the resulting current subtree S''' has broadcast time at most t, and thus it is the only component in its optimal partitioning, $b(S''', t) = 1$. Thus $c(T, S''') + b(S''', t) = b(T, t)$, and as no more cuts are made, $c(T, \Phi) = c(T, S''') = b(T, t) - 1$. □

The pruning strategy employed in the algorithm BROADCAST guarantees that each vertex is processed exactly once, and thus the complexity of the algorithm is defined by the complexity of upmaxfather and upminfather, which are executed at most once per processed vertex. These procedures, in turn, involve at most one call of emptyslots and/or insert which require number of operations in the order of length of the relevant (callee) list. This list of light sons of a vertex in u is never longer than t. Hence, we have the following theorem determining the complexity of BROADCAST.

THEOREM 2. *Given a tree T with n vertices and a broadcast time t, the execution time of algorithm* BROADCAST *is $O(n \cdot t)$.*

4. Conclusion. In this paper, we have presented a linear algorithm for decomposing a given tree into subtrees, each subtree having a broadcast time less than or equal to a given time. This algorithm can be seen as one of a family of linear tree partitioning algorithms [4]. Partitioning techniques can be seen as alternatives to methods determining multiple centers (cf. [7] and [4]).

Partitioning based on broadcast time is well-motivated from an applications perspective. Other models of the information dissemination process would lead to different decomposition problems. For example, associating a call time with each line to reflect average load (i.e., queue length) is a reasonable extension of our model.

REFERENCES

[1] A. M. FARLEY, *Minimal broadcast networks*, Networks 9 (1979), 313–332.
[2] A. M. FARLEY, *Broadcast time in communication networks*, SIAM J. Applied Math., 39 (1980), pp. 385–390.
[3] A. M. FARLEY, S. T. HEDETNIEMI, S. L. MITCHELL AND A. PROSKUROWSKI, *Minimum broadcast graphs*, Discr. Math, 25 (1979), pp. 189–193.
[4] A. M. FARLEY, S. T. HEDETNIEMI AND A. PROSKUROWSKI, *Partitioning trees: matching, domination and maximum diameter*, Internat. J. Comp. Inform. Sci., to appear.
[5] M. R. GAREY AND D. B. JOHNSON, *Computers and Interactability*, W. H. Freeman, San Francisco, 1978.
[6] S. M. HEDETNIEMI AND S. T. HEDETNIEMI, *Broadcasting by decomposing trees into paths of bounded length*, CS-TR-79-16, University of Oregon, Eugene, OR, 1979.
[7] O. KARIV AND S. L. HAKIMI, *Algorithmic approach to network location problems* I: *the p-centers*, SIAM J. Appl. Math, 37 (1979), pp. 513–530.
[8] D. E. KNUTH, *The Art of Computer Programming*, vol. I, 2nd ed., Addison-Wesley, Reading, MA, 1973.
[9] A. PROSKUROWSKI, *Minimum broadcast trees*, IEEE Trans. Computers, C-30 (1981), pp. 363–366.
[10] P. J. SLATER, E. J. COCKAYNE AND S. T. HEDETNIEMI, *Information dissemination in trees*, SIAM J. Comput., 10 (1981), pp. 692–701.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 4, December 1981

0196-5212/81/0204-0004 $01.00/0

THE BANDWIDTH OF CATERPILLARS WITH HAIRS OF LENGTH 1 AND 2*

S. F. ASSMANN†, G. W. PECK†, M. M. SYSŁO‡ AND J. ZAK§

Abstract. In this paper we show that the bandwidth of any caterpillar with hairs of length 1 and 2 is given by the maximum over all subcaterpillars of $\lceil (n-1)/d \rceil$, where n is the number of vertices and d is the diameter of the subcaterpillar. We also give an $n \log n$ algorithm which produces a bandwidth labelling of such a caterpillar.

Let G be a connected graph with vertex set V and edge set E. A *labelling* of G is a 1–1 map f from V into N, the nonnegative integers. Let $b_f(C) = \max |f(u) - f(v)|$, where (u, v) ranges over all edges of G. The *bandwidth* of G, denoted $b(G)$, is defined by $b(G) = \min b_f(G)$, where f ranges over all labellings of G. A labelling f such that $b_f(G) = b(G)$ is called a *bandwidth labelling* of G. The paper of Chinn et al. [1] is a general survey of bandwidth theory and its uses.

Finding the bandwidth of a graph has several practical applications. Suppose the edges of G correspond to the nonzero entries in a symmetric $n \times n$ matrix M. That is, there is an edge from vertex i to vertex j if and only if M_{ij} does not equal 0. Finding a bandwidth labelling of G then corresponds to permuting the rows and columns of M in such a way as to minimize the maximum distance from any nonzero entry of M to the main diagonal. This distance will equal $b(G)$.

Another application is the following. Suppose we wish to carry out some iterative procedure on the vertices of a graph, where the new value associated with a vertex is determined by the old values associated with its neighbors and itself. Then the maximum number of old values we must keep in memory at one time will be $2b(G)+1$, and we can achieve this if we process the vertices in the order given by the bandwidth labelling.

Finding the bandwidth of an arbitrary connected graph is an NP-complete problem [4]. In fact, the problem is NP-complete even if the graphs are restricted to trees with maximum degree 3 [3]. However, Chung [2] gives formulas for the bandwidth of a few special classes of graphs, such as the n-cube and the complete p-ary tree with k levels.

In this paper we give the bandwidth for another special class, the caterpillars with hairs of length 1 and 2. That this bandwidth can be found in $n \log n$ time is interesting, since the trees used in [3] to prove NP-completeness look very much like such graphs.

A *caterpillar* is a graph in which the removal of all pendant vertices results in a path. These pendant vertices can be thought of as *hairs* attached to the *body* of the caterpillar, that is, the path of nonpendant vertices. See Fig 1. We will also consider *caterpillars with hairs of length* 1 *and* 2. In these graphs, paths of length 1, 2 or both may be attached to any vertex in the body of the caterpillar. See Fig. 2. In these two figures the bodies of the caterpillars are shown as black dots, the hairs as white dots.

One can easily show that inequality (1) holds, where G' ranges over all connected subgraphs of G, n' is the number of vertices of G', and d' is the diameter of G'.

$$\max \lceil (n'-1)/d' \rceil \leqq b(G). \tag{1}$$

* Received by the editors September 8, 1980, and in revised form March 30, 1981.

† Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

‡ Computer Science Department, Washington State University, Pullman, Washington 99164. Permanent address: Institute of Computer Science, University of Wrocław, Pl. Grunwaldzki 2/4, 50384 Wrocław, Poland.

§ K. Jadwigi 13, 48200 Prudnik, Poland.

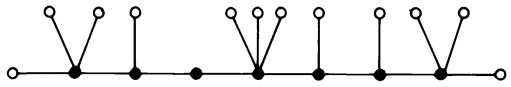


FIG. 1. *A caterpillar.*

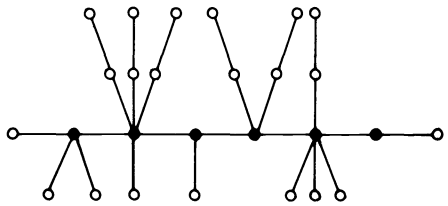


FIG. 2. *A caterpillar with hairs of length* 1 *and* 2.

The inequality in (1) cannot in general be strengthened to an equality, as the example in Fig. 3 shows. In this graph, the lower bound is 2, but $b(G) = 3$.

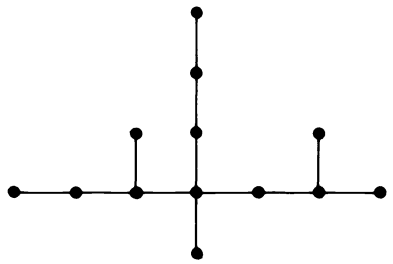


FIG. 3. *A counterexample to equality in* (1).

We will now show that (1) *is* strengthened to an equality when G is a caterpillar with hairs of length 1 and 2.

We present an algorithm which, given a graph G of this type and a positive integer m, attempts to find a labelling f and G such that $b_f(G) = m$. We show that if the algorithm fails to produce such a labelling, then there is a subcaterpillar G' of G with $\lceil (n'-1)/d' \rceil > m$, so that by (1) we have $b(G) > m$.

The algorithm is as follows:

ALGORITHM 1.

Step 1. Label the points along a diameter of G by 0, m $2m, \cdots, dm$ in order from left to right.

Step 2. For $k = 1$ to $d-1$ label the hairs of point km in the following order, giving each vertex the lowest possible label consistent with preserving bandwidth m. If a point cannot be labelled, halt and return an error message.

Step 2A. As many hairs of length 2 as can be labelled so that the point further from the point km is given a label between $(k-2)m$ and $(k-1)m$.

Step 2B. All hairs of length 1.

Step 2C. As many hairs of length 2 as can be labelled so that both points receive labels between $(k-1)m$ and km.

Step 2D. All remaining points at distance 1 from point km.

Step 2E. All remaining points at distance 2 from point km, taken in the same order as their associated points in step 2D.

As as example, the labelling the algorithm produces for the graph in Fig. 2 when $m = 4$ is given in Fig. 4.

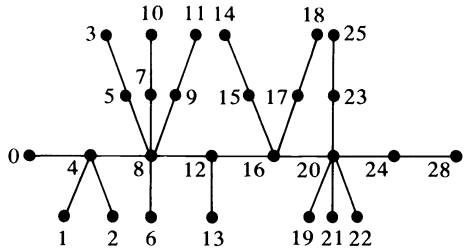


FIG. 4. *An example of the use of Algorithm* 1.

The idea behind the algorithm is that no point at distance 2 from point km is given a label between $(k-1)m$ and $(k+1)m$ unless, as in step 2E, there are no points at distance 1 which need labels, or, as in step 2C, it is optimal to do so. (If we do not use two labels within m of km to label this hair, we must use one between km and $(k+1)m$ and one above $(k+1)m$, wasting the one below km.)

(*) Note that, when the hairs of point km, are labelled, no label between km and $(k+1)m$ is assigned unless all labels between $(k-1)m$ and km have been used, and no label between $(k+1)m$ and $(k+2)m$ is assigned unless all labels between km and $(k+1)m$ have been used.

(**) Also note that each interval jm to $(j+1)m$ is used in order from lowest to highest label.

THEOREM 1. *Let G be a caterpillar with hairs of length* 1 *and* 2, *and let m be a positive integer. If Algorithm* 1 *fails to find a labelling f of G with $b_f(G) = m$, then $b(G) > m$.*

Proof. Suppose the algorithm fails. We will then show that we can find a subcaterpillar G' of G with $\lceil (n'-1)/d' \rceil > m$. By (1) we will then have $b(G) > m$.

The general idea of the proof is as follows. Let km be the point whose hairs we have failed to finish labelling. Suppose that all the labels below km have been used. Then (with a few exceptions) G' will consist of all points with labels below km and all points no further from km than the point we have failed to label. In the first part of the proof we show that we can reduce any problem to the special case where all the labels below km have been used by deleting all points with labels below the gap. In the second part of the proof we explain in more detail how to obtain G' and why $\lceil (n'-1)/d' \rceil$ will be greater than m.

First, suppose that some labels less than km have not been used. Let i be the greatest such label. We will show that the algorithm would also fail on the subcaterpillar G'' obtained from G by deleting all points with labels less than i.

If $i < (k-2)m$, then any points with labels less than i are too far from the point km to have any effect on the labelling of its hairs, so we can delete them to form G'' without making it possible to label the unlabelled hair.

If $(k-2)m < i < (k-1)m$, then by remark (*) no hairs from point $(k-2)m$ or point $(k-1)m$ can have labels between $(k-1)m$ and km. Before we started step 2 for km, then, there were $m-1$ unused labels between $(k-1)m$ and km and at least 1 and no

more than $m-1$ unused labels between $(k-2)m$ and $(k-1)m$. Since there is still a gap between $(k-2)m$ and $(k-1)m$, point km must have no unlabelled hairs of length 2, or we could have continued to apply step 2A. Furthermore, all points on the hairs of length 2 which are at distance 2 from point km have labels below i, by (**). The gap is too far away to affect the labelling of points at distance 1 from point km. Thus, the algorithm would also fail on G''.

It is impossible to have $(k-1)m < i$ because the algorithm will not fail until all labels within m of km have been used.

By remark (**), $i = jm - 1$ for some j, so the partial labelling Algorithm 1 would give G'' is the same as it gave G'' as a subgraph of G except for a shift of jm. Thus the labelling of G'' would have no gaps.

We may therefore assume for the rest of the proof that there are no gaps in the partial labelling that Algorithm 1 gives G before it fails.

G' will depend on which step in the algorithm we fail in.

It is impossible for the algorithm to fail in step 2A because we have not exhausted the possible labellings of hairs of length 2.

Suppose it fails in step 2B. There are 2 cases.

Case 1. We have not labelled any hairs in step 2A. Let G' consist of all points with labels no more than km and all points at distance 1 from point km. G' has diameter $k+1$. G' includes the $(k+1)m+1$ points labelled 0 through $(k+1)m$, plus at least 1 more point which we could not label. Thus $n' \geqq (k+1)m+2$, so $\lceil (n'-1)/d' \rceil > m$. See Fig. 5 for an example of this situation.

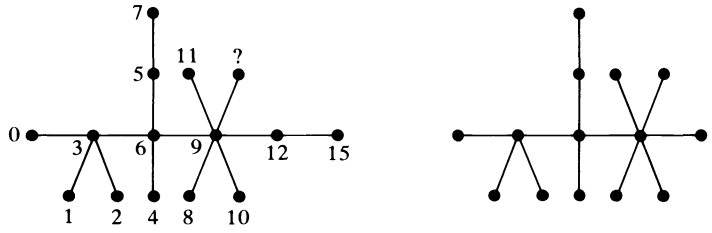


FIG. 5. (a) *Failure in step* 2B, *where step* 2A *has not been used.* (b) G'. *Note* $\lceil (14-1)/4 \rceil = 4$.

Case 2. We have labelled at least one hair in step 2A. Let G' consist of the point km and all points at distance 1 from point km. Because there had been a label free between $(k-2)m$ and $(k-1)m$ which allowed us to use step 2A, by remark (*) no labels above $(k-1)m$ were used to label any hairs of points below km. Since we are stuck in step 2B, all the labels from $(k-1)m$ to $(k+1)m$ must have been given to points at distance 1 from point km, except for the label km itself. We also have at least one more point at distance 1 from km which we could not label. Thus $n' \geqq 2m+2$. As $d'=2$, $\lceil (n'-1)/d' \rceil > m$. See Fig. 6 for an example of this situation.

We cannot fail in step 2C, because we still have not exhausted the possible labellings of hairs of length 2.

Suppose we fail in step 2D. There are again 2 cases.

Case 1. We did not label any hairs in step 2C. Then we are in essentially the same situation as if we had failed in step 2B, with the same 2 subcases. See Figs. 7 and 8 for examples of these situations.

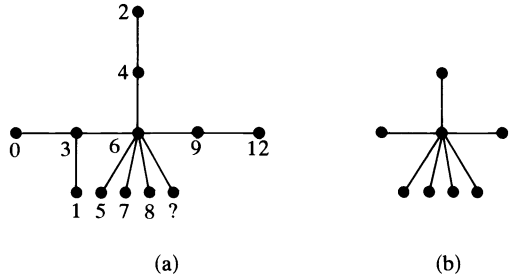


FIG. 6. (a) *Failure in step* 2B *where step* 2A *has been used.* (b) G'. *Note* $\lceil (8-1)/2 \rceil = 4$.

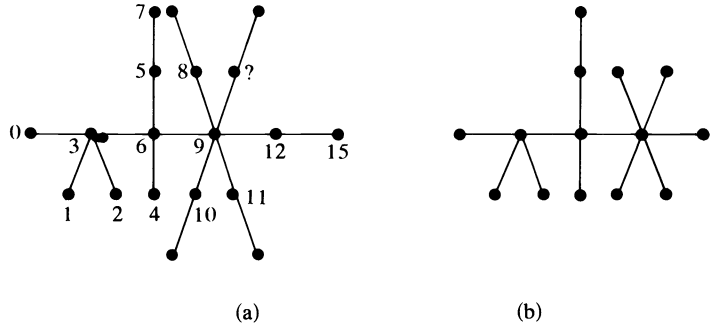


FIG. 7. (a) *Failure in step* 2D, *where step* 2C *has not been used, and step* 2A *has also not been used.* (b) G'. *Note* $\lceil (14-1)/4 \rceil = 4$.

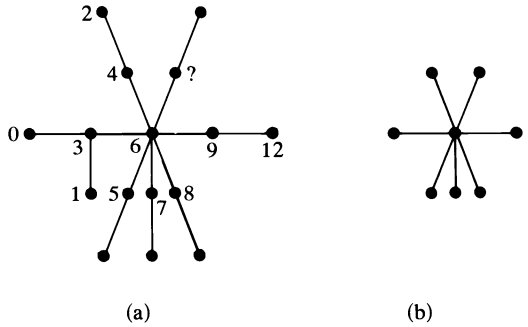


FIG. 8. (a) *Failure in step* 2D, *where step* 2C *has not been used, but step* 2A *has been used.* (b) G'. *Note* $\lceil (8-1)/2 \rceil = 4$.

Case 2. We did label at least 1 hair in step 2C. Let G' consist of all points with labels no more than km and all points at distance 1 or 2 from point km. Because we have failed in this step, G' includes all the points labelled 0 through $(k+1)m$, the point labelled $(k+2)m$, and the point we could not label. Also, at least the points labelled $km+1$ to $(k+1)m-1$ and the unlabelled point each have associated with them a point at distance 2 from point km which hasn't been labelled yet. Thus $n' \geqq (k+2)m+3$. The diameter $d' = k+2$, so $\lceil (n'-1)/d' \rceil > m$. See Fig. 9 for an example of this situation.

Suppose we fail in step 2E. Let G' consist of all points with labels no more than km and all points at distance 1 or 2 from point km. If every point labelled in step 2D

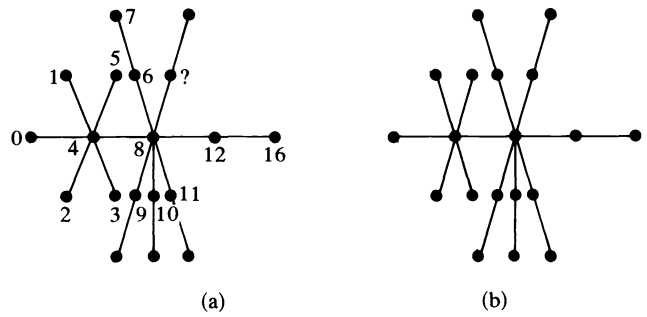


FIG. 9. (a) *Failure in step* 2D, *where step* 2C *has been used.* (b) G'. *Note* $\lceil (19-1)/4 \rceil = 5$.

was given a label above km, then its associated point could be given the label m above that label in step 2E, and we would not be stuck. Thus one point in step 2D was given the label $km-1$, and the rest were given the labels $km+1$ to $(k+1)m-1$. G' includes all the points with labels 0 through $(k+1)m$, the point $(k+2)m$, and m points at distance 2 from point km which have not yet been labelled. Thus $n' = (k+2)m+2$. The diameter $d' = k+2$, so $\lceil (n'-1)/d' \rceil > m$. See Fig. 10 for an example of this situation.

This concludes the proof of the theorem.

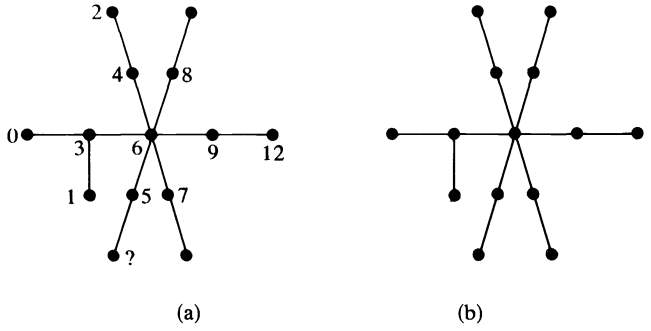


FIG. 10. (a) *Failure in step* 2E. (b) G'. *Note* $\lceil (14-1)/4 \rceil = 4$.

THEOREM 2. *The bandwidth of a caterpillar G with hairs of length* 1 *and* 2 *is the maximum over all subcaterpillars G' of* $\lceil (n'-1)/d' \rceil$.

Proof. Apply Algorithm 1 with $m = \max \lceil (n'-1)/d' \rceil$. The algorithm cannot fail, so $b(G) \leqq m$. Now use (1).

COROLLARY. *There is an n log n algorithm for finding the bandwidth of a caterpillar G with an points and hairs of length* 1 *and* 2, *which also produces a labelling achieving this bandwidth.*

Proof. Clearly the bandwidth will be between 1 and n. We can find $b(G)$ by binary search in this interval, using Algorithm 1 to test whether $b(G) > m$. We will have to use Algorithm 1 log n times. Algorithm 1 itself works in time proportional to the number of points in G, so the total time used is n log n.

Note that once we have any bandwidth labelling of a graph, it can easily be changed to a bandwidth labelling with range $\{1, 2, \cdots, n\}$ by ordering the vertices by label and letting the new label of each vertex be its position in this ordering.

The example of Fig. 3 showed that Theorem 2 is false for caterpillars with hairs of length 1, 2, and 3. Figure 11 shows that the theorem is false for caterpillars which have all their hairs of length 4. Here the lower bound is 2 but $b(G) = 3$.

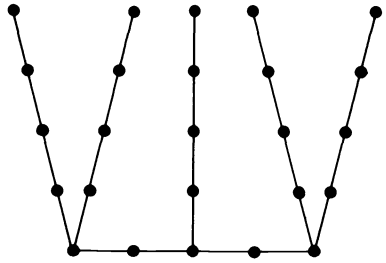


FIG 11. *A counterexample to equality in* (1) *where all hairs have the same length.*

The following question remains open. Does there exist an efficient algorithm, perhaps similar to Algorithm 1, which will determine whether a caterpillar with hairs of length no more than k has a bandwidth below a given bound?

Another open question is how bad the lower bound $\max \lceil (n'-1)/d' \rceil$ can be in comparison to the actual bandwidth of a graph.

Acknowledgments. One of the authors (GWP) thanks Jeff Kahn, Daniel Kleitman, and Jim Shearer for their help in providing his contribution to the paper.

REFERENCES

[1] P. Z. CHINN, J. CHVATALOVA, A. K. DEWDNEY AND N. E. GIBBS, *Graph bandwidth: a survey of theory and applications*, J. Graph Theory, submitted.

[2] F. R. K. CHUNG, *Some problems and results on labellings of graphs*. Proc. 4th International Conference on the Theory and Applications of Graphs, Kalamazoo, John Wiley, New York, 1980.

[3] M. R. GAREY, R. L. GRAHAM, D. S. JOHNSON AND D. E. KNUTH, *Complexity results for bandwith minimization*, SIAM J. Appl. Math., 34 (1978), pp. 477–495.

[4] C. H. PAPADIMITRIOU, *The NP-completeness of the bandwith minimization problem*, Computing, 16 (1976), pp. 263–270.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 4, December 1981

0196-5212/81/0204-0005 $01.00/0

COVERING REGIONS BY RECTANGLES

SETH CHAIKEN†, DANIEL J. KLEITMAN‡, MICHAEL SAKS§ AND JAMES SHEARER¶

Abstract. A *board* $\mathcal{B}$ is a finite set of unit squares lying in the plane whose corners have integer coordinates. A *rectangle* of $\mathcal{B}$ is a rectangular subset of $\mathcal{B}$ and an *antirectangle* is a set of squares in $\mathcal{B}$ no two of which are in a common rectangle. We prove a conjecture of Chvátal that if $\mathcal{B}$ is convex in the horizontal and vertical directions, then the minimum number of rectangles whose union is $\mathcal{B}$ equals the maximum cardinality of an antirectangle. Our proof uses two analogous minimax theorems about covering the corners and covering the edges of the board.

We quote examples that illustrate the necessity of the hypotheses, and give some conjectures and open questions. The method of proof can give a polynomial running time algorithm for finding a minimum cover.

1. Introduction. Consider the plane covered by the unit squares whose sides lie on the integer coordinate lines; we refer to these throughout as *squares*. A *board* $\mathcal{B}$ of size n is a (finite) set of n squares. A *rectangle* (in $\mathcal{B}$ unless otherwise indicated) is a subset of $\mathcal{B}$ whose union is rectangular. A *whole cover* of $\mathcal{B}$ is a collection of rectangles whose union equals $\mathcal{B}$. The rectangles of a cover may overlap, but each of them must be wholly contained in the board. An *antirectangle* in $\mathcal{B}$ is a set of squares in $\mathcal{B}$ no two of which are contained in any rectangle. Any cover must contain at least as many rectangles as any antirectangle has squares. Therefore, if θ is the number of rectangles in a cover (the size of the cover) and α is the number of squares in an antirectangle (the size of the antirectangle) then $\theta \geqq \alpha$. We call a cover optimal if it has minimum size and an antirectangle optimal if it has maximum size. If a board has a cover and an antirectangle of equal size, then they are both optimal.

The problem of finding optimal covers and antirectangles is an example of a dual pair of packing and covering problems, well known in combinatorics (see, for example, Liu [5], Brualdi [2], Woodall [6] for more details). Chvátal originally conjectured that the optimal θ and α were equal. In general, this is false (see § 3). Here we prove his weakened conjecture that there is equality when $\mathcal{B}$ is *convex*: Whenever two squares in $\mathcal{B}$ are on the same horizontal or vertical line, all squares between them are in $\mathcal{B}$. This problem arose as an idealized special case of an operation used by the microelectronics industry. A layer of an integrated circuit (consisting of arbitrary polygons) is to be printed on a photographic plate that will become a photolithographic mask in the manufacture of the integrated circuit. The printing is done by flashing rectangles onto the photographic plate to produce an image equal to their superposition; this is to be done using as few rectangles as possible. In the "real world problem" there are additional constraints, including the discreteness of the rectangles available (which limits the accuracy), not exposing a segment of a polygon boundary more than once, and computation time and program and data space limitations. In retrospect, the

* Received by the editors February 22, 1980, and in final form March 12, 1981. This research was supported in part by the Office of Naval Research under contract N00014-76-C-0366.

† Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. Present address: Department of Mathematics, Department of Computer Science, State University of New York at Albany, Albany, New York 12222.

‡ Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

§ Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. Present address: Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903.

¶ Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.
The work of this author was supported in part by a National Science Foundation Fellowship.

theory here in part supports certain heuristics that have been used in such a program.[1] Another context (Masek [5]) is the construction of letters and other shapes on video computer terminals.

Our method of proof can be used to obtain a polynomial time algorithm for finding the optimal θ, but we omit the details. Masek [5] established that, in general, for nonconvex boards, this problem is NP-hard. This is yet another example of a combinatorial optimization problem with a min-max theorem and an efficient algorithm, and a problem without such a theorem that is NP complete.

For any subset $S \subset \mathcal{B}$, an S-cover of $\mathcal{B}$ is a collection of rectangles whose union contains S. An S-antirectangle of $\mathcal{B}$ is an antirectangle contained in S. Let $\mathcal{E} = \mathcal{E}(\mathcal{B})$ be the set of *edge* squares, that is, those with at least one side lying on the boundary of $\mathcal{B}$. Let $\mathcal{C} = \mathcal{C}(\mathcal{B})$ be the set of *corner* squares, those with two adjacent sides on the boundary. Nonedge squares are called *interior* squares. Our proof of the main theorem relies on induction for certain boards (called *reducible*). For the remaining (*irreducible*) boards, the proof uses an analogous min-max result for edge covers and edge antirectangles, which holds for these boards. The proof of this edge result makes use of a theorem about corner covers and antirectangles.

These problems can be restated in familar graph theoretic terminology by associating a board $\mathcal{B}$ with a graph $G = G(\mathcal{B})$ whose nodes are the squares in $\mathcal{B}$ and in which two squares are joined by an arc if there is a rectangle in $\mathcal{B}$ that contains them both. The following simple lemma, which is true for any board, is presented without proof.

LEMMA 1.1 *The cliques of $G(\mathcal{B})$ are the rectangles of $\mathcal{B}$.*

Let S be any subset of $\mathcal{B}$ and let G_S denote the induced subgraph of G on this subset. A maximum S-antirectangle of $\mathcal{B}$ is, by definition, a maximum independent set of G_S, whose size is written $\alpha(G_S)$. A minimum S-cover of $\mathcal{B}$ has size equal to $\theta(G_S)$, the minimum number of cliques needed to cover G_S.

Our main results, which hold for convex boards $\mathcal{B}$, are:

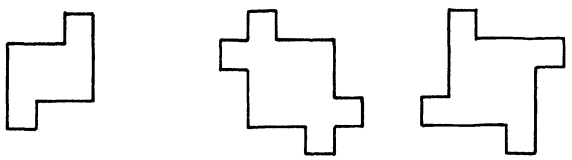
THEOREM I. *The minimum size of a corner cover of $\mathcal{B}$ equals the maximum size of a corner antirectangle in $\mathcal{B}$; i.e., $\theta(G_{\mathcal{C}}) = \alpha(G_{\mathcal{C}})$.*

LEMMA. 6.5. *If $\mathcal{B}$ is irreducible (see § 5), then the minimum size of an edge cover of $\mathcal{B}$ equals the maximum size of an edge antirectangle in $\mathcal{B}$, i.e., $\theta(G_{\mathcal{E}}) = \alpha(G_{\mathcal{E}})$.*

THEOREM II. *The minimum size of a whole cover of $\mathcal{B}$ equals the maximum size of an antirectangle in $\mathcal{B}$, i.e., $\theta(G) = \alpha(G)$.*

It is shown (§ 3) that the convexity hypothesis in these results is necessary. Figure 1 illustrates that the sizes of the optimal corner, edge, and whole covers may be different.

Theorem I is proven by showing that $G_{\mathcal{C}}$ has a simple structure.



No optimal corner cover covers all the edges.

No optimal edge cover covers the whole board.

FIG. 1

[1] The first author encountered this problem during his employment with Applicon, Inc.

LEMMA 4.1. *Each connected component of* $G_{\mathscr{C}}$ *is either*

(1) *a* 4 *clique, or*

(2) *a graph in which every odd cycle contains a square of degree* 2, *whose neighbours are adjacent.*

To prove Lemma 6.5, we first show that irreducibility implies that in a corner cover by maximal rectangles, the rectangle covering a corner covers all of the squares on at least one of the edges incident to the corner. This fact enables us to construct an arc-deleted subgraph G^* of $G_{\mathscr{C}}$ which satisfies $\alpha(G^*) = \theta(G^*)$ and whose independent sets and clique covers correspond to $\mathscr{E}$-antirectangles and $\mathscr{E}$-covers.

Theorem II is proved in two steps. First, two *reducible configurations* are defined. Each configuration involves a maximal rectangle that must be in every optimal cover, and a reduction which produces a smaller board. The reduction is such that from an equally sized cover and antirectangle pair for the reduced board (obtained by induction), we can construct an equally sized pair for the original board. The second step involves analysis of irreducible boards. If some optimal edge cover covers the whole board, Lemma 6.5 implies the result. If not, we show the board is so structured that an optimal edge cover that covers a maximal set of squares provides an equally sized antirectangle. To these, one more rectangle and one more square can be added to yield an optimal whole cover and antirectangle.

In the next section are definitions and simple facts used in the rest of the paper. Section 3 gives examples that indicate the necessity of the convexity hypothesis, and some open problems. The remainder of the paper is restricted to convex boards. Section 4 is about corner covering. Section 5 gives the reducible configurations. Section 6 presents the edge covering result. Finally, in § 7 we finish covering the whole board when it is irreducible.

2. Definitions and simple facts. We adopt the usual coordinate system in which the positive x axis points to the *right* and the y axis points *up*. The defines for us the everyday words for direction: top, below, horizontal, etc. The squares are the unit squares bounded by integer coordinate lines. We identify a board or rectangle (set of squares) with its union (a polygon). A board has *vertices* and *edges* on its boundary, but the edges of a rectangle are called *sides.*

A vertex with interior included angle of 90° is called a *corner vertex*; a *corner square* is one that touches a corner vertex. The integer coordinates divide each edge into unit (boundary) *segments.* A square that is bounded on at least one side by a segment is called an *edge square.* Every corner vertex is associated to a unique corner square and every segment is associated to a unique edge square. Let $\mathscr{C}$ and $\mathscr{E}$ denote, respectively, the sets of corner and edge squares.

Points, such as vertices, are referred to by their coordinates (x, y). We make the special convention that the cordinates of a square or segment are those of its center. The coordinates of z are denoted by $x(z)$ and $y(z)$. When z is an edge, $t \geqq x(z)$ means $t \geqq x(p)$ for all points p in z. This way, for example, we say square u is to the right of vertical edge CD and left of vertical edge EF by $x(CD) < x(u) < x(EF)$. Occasionally, $x \pm 0.5$ is used to convert between coordinates of points and those of squares or segments.

A rectangle is said to *cover* a corner, a square or an edge if it contains the corresponding corner square, the indicated square, or all the edge squares on the edge respectively.

An edge is called a *support edge* when both its ends are corner vertices. It is easy to see every support edge has a unique maximal rectangle that covers it. This is called

the *associated* or *support edge rectangle.* Any rectangle can be named by giving either a point, segment, or square on two opposite corners. Thus $\square uv$ denotes the smallest rectangle that contains u and v.

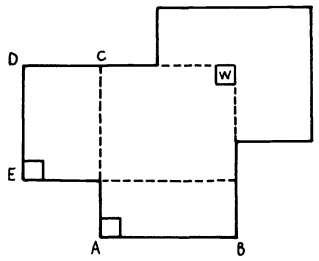


FIG 2. $\square BC$ *is the support edge rectangle of* AB; $\square EW$ *is the support edge rectangle of* DE.

$G = G(\mathcal{B})$ is a graph whose nodes, called squares, are the squares in $\mathcal{B}$. Two squares s and t are joined by an arc, denoted $(s, t) \in G$, whenever $\square st \subseteq \mathcal{B}$. An easy way to check whether $(s, t) \notin G$ is to check whether the interior of $\square st$ meets a boundary segment or a square not in $\mathcal{B}$. The neighborhood of a square s, $N_G(s) = N(s)$, is the set of all squares that can be covered by some rectangle covering s:

$$N_G(s) = \{t \in \mathcal{B} \mid (s, t) \in G\} \cup \{s\}$$

Let R be the associated rectangle of support edge AB. The side of R opposite AB must contain a boundary segment. Let e be the edge square on AB that meets the perpendicular bisector of that segment. R is the unique maximal rectangle that covers e. Furthermore, in any cover and antirectangle problem in which e must be covered, there is always an optimal cover that contains R and an optimal antirectangle that contains e.

All of our positive results concern *convex* boards as defined in § 1. In other words, for any $s_1, s_2 \in \mathcal{B}$, if $x(s_1) = x(s_2)$ or $y(s_1) = y(s_2)$ then $\square s_1 s_2 \subseteq \mathcal{B}$. We make repeated use of the following facts about convex $\mathcal{B}$:

Fact 2.1. Given a pair of squares, s_1, $s_2 \in \mathcal{B}$, consider the other two squares at the corners of $\square s_1 s_2$, $s_1' = (x(s_1), y(s_2))$, $s_2' = (x(s_2), y(s_1))$. When (s_1, s_2) is in G, then s_1' and s_2' are in $\mathcal{B}$. Convexity means that the converse is true. Then, we need to look at only two squares, s_1' and s_2', to see whether $(s_1, s_2) \in G$.

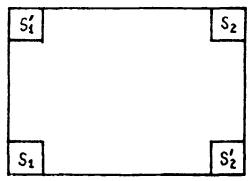


FIG. 3. *Sufficient condition for* $(s_1, s_2) \in G$ *for convex* $\mathcal{B}$.

Fact 2.2. A convex board has exactly 4 support edges. These divide the boundary into 4 support edges and 4 possibly empty paths.

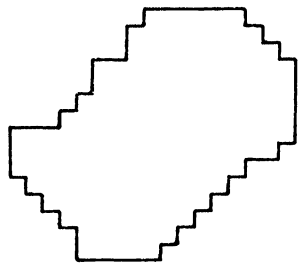


FIG. 4. *Note the four support edges of a convex board.*

The arguments made in succeeding sections hold under rotations and reflections of the board. For simplicity, the arguments are described and illustrated with the board in a specific position.

3. Counterexamples and open questions. Chvátal's original conjecture was disproved by Szemerédi who found a counterexample with a "hole" (Fig. 5). Chung (who informed us of the history of this problem) then found the simply connected counterexample in Fig. 6 (Chung [3]).

One can see that optimal θ and α in these examples are unequal by first observing that a support edge rectangle R always contains some edge square such that R is the unique maximal rectangle that contains that square. Thus one can assume that the support edge rectangles are all in the optimal cover and that one edge square from each is in the optimal antirectangle. Second, consider the squares S left uncovered by these rectangles (not cross hatched in the drawings). In each example, G_S has an induced 5-cycle (indicated by connected dots in the drawings). Hence at least 3 cliques are needed to cover S. One can verify that 3 cliques suffice. Finally, one can verify there are only up to two independent squares in S.

Similar analyses of Figs. 7 and 8 yield 7-cycles, and so show that the corner covering result ($\theta(G_{\mathscr{C}}) = \alpha(G_{\mathscr{C}})$) and the edge covering result ($\theta(G_{\mathscr{E}}) = \alpha(G_{\mathscr{E}})$) are sometimes false for nonconvex boards.

A graph G is perfect if for all subsets S of vertices, $\theta(G_S) = \alpha(G_S)$ (see Berge [1]). Figure 9 shows a board with a subset of squares that includes a 5-cycle in G; hence $G(\mathscr{B})$ is not always perfect, even for convex boards.

In this paper we show that $\theta(G_{\mathscr{E}}) = \alpha(G_{\mathscr{E}})$ for certain (irreducible) convex boards (§ 6). One of the authors, M. Saks, has recently shown that for any convex board, and any subset $\mathscr{E}'$ of edge squares, $\theta(G_{\mathscr{E}'}) = \alpha(G_{\mathscr{E}'})$; hence $G_{\mathscr{E}}$ is a perfect graph.

We have not yet fully investigated the implications of convexity in only one direction.

Finally, for arbitrary boards $\mathscr{B}$, let θ and α be the optimal cover and antirectangle sizes respectively. Erdös asked if θ/α is bounded and we do not know the answer. Chung's example has $\theta/\alpha = \frac{8}{7}$. The largest value we could achieve for θ/α is $\frac{21}{17} - \varepsilon$,

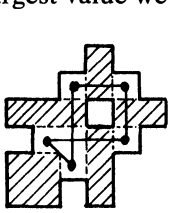


FIG. 5. $\theta(G) = 8$, $\alpha(G) = 7$.

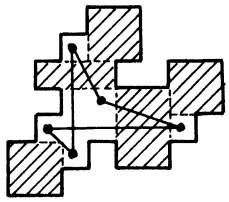


FIG. 6. $\theta(G)=8$, $\alpha(G)=7$.

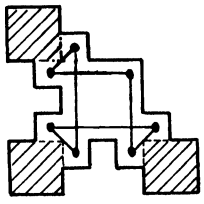


FIG. 7. $\theta(G_{\mathscr{C}})=7$, $\alpha(G_{\mathscr{C}})=6$.

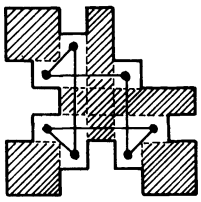


FIG. 8. $\theta(G_{\mathscr{C}})=9$, $\alpha(G_{\mathscr{C}})=8$.

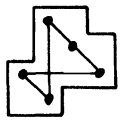


FIG. 9. *Induced 5-cycle in G.*

4. Corner covering. A corner vertex of the board is designated as type (left, upper), (left, lower), (right, upper), or (right, lower) according to its position with respect to the board square it touches. A corner square has the types of its incident corner vertices (it may have one or two for nontrivial boards). The types of two corners c_1, c_2 provide necessary conditions for $(c_1, c_2) \in G$.

If c_1, c_2 have a common type, $(c_1, c_2) \notin G$ (Fig. 11).

If c_1, c_2 each have one type, and the types differ in both components, we say c_1, c_2 have opposite types. Whether $(c_1, c_2) \in G$ or not depends on the rest of $\mathscr{B}$ (Fig. 12).

If c_1, c_2 have types that differ in one component (say the x, i.e., "left", "right" component) and agree in the other, we say they have adjacent type. In this case, $(c_1, c_2) \in G$ only if c_1 and c_2 have equal coordinates in the component in which their types agree (say the y component). This condition is sufficient when $\mathscr{B}$ is convex (Fig. 13).

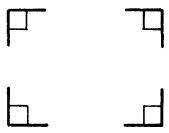


FIG. 10. *The four corner types.*

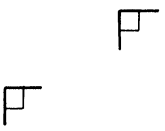


FIG. 11. *Corners with common type.*

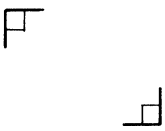


FIG. 12. *Corners with opposite type.*

FIG. 13. *Corners with adjacent type.*

LEMMA 4.1. *For a convex board β, each connected component of $G_{\mathscr{C}}$ is either*

(1) *a 4 clique, or*

(2) *a graph in which every odd cycle contains a square of degree 2 in $G_{\mathscr{C}}$ such that its two neighbors are adjacent (thus these three vertices from a 3 clique).*

Proof. Let $(c_1, \cdots, c_n)$, $n \geqq 3$, be a odd cycle in $G_{\mathscr{C}}(\mathscr{B})$. Then at least two successive squares in the cycle, say c_1 and c_2, are of adjacent type.

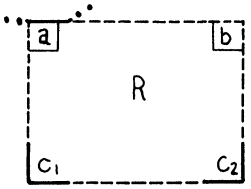


FIG. 14

There is a unique maximal rectangle R that contains both c_1 and c_2. Convexity implies the side of R opposite c_1 and c_2 meets a boundary segment at at least one of its ends; assume that end is closest to c_1 (Fig. 14). Therefore, $R = N_G(c_1)$. Thus in $G_{\mathscr{C}}$, c_1 is connected only to c_2 and to whichever of a and b are corners. If both a and b are corners, c_1, c_2, a, b is a 4 clique that is not connected to any other corner (Fig. 10). If only one of a and b, say a, is a corner, then c_1 has degree 2 and a, c_1, c_2 is a 3-clique. If neither is a corner, then c_1 is not contained in a cycle.

THEOREM I. *The minimum size of a corner cover of $\mathscr{B}$ equals the maximum size of a corner antirectangle in $\mathscr{B}$.*

Proof of Theorem I. We show that any graph satisfying the properties proved in Lemma 4.1 has a clique cover and independent set of equal size. The cover consists of all of the (disconnected) 4-cliques, all 3-cliques, and a minimum cover of the subgraph H obtained by deleting all of these cliques. The independent set consists of one square from each 4-clique, a degree 2 square from each 3-clique and a maximum independent set in H. (This set is independent since every square in H is independent of the degree 2 squares in each 3-clique.) Now H is bipartite, i.e., it has no odd cycles, since otherwise H would contain a 3-clique by Lemma 4.1. It is a well-known consequence of the König–Egerváry theorem that the size of the minimum clique cover of a bipartite graph equals that of the largest independent set. Thus the given cover and independent set of $G_{\mathscr{C}}$ have the same size. □

In subsequent sections, we will need the following stronger result.

LEMMA 4.2. *Say H is a subgraph of $G_{\mathscr{C}}$ obtained by deleting some arcs such that whenever the arc joining the two neighbors of a degree 2 square s of a 3-clique is deleted, an arc incident to s is also deleted. Then $\alpha(H) = \theta(H)$.*

Proof. It is easy to see that deleting any such arcs preserves the property of $G_{\mathscr{C}}$ proved in Lemma 4.1 and thus the proof of Theorem I applies.

5. Reducible configurations. Assume $\mathscr{B}$ is a convex board.

THEOREM II. *The minimum size of a whole cover of $\mathscr{B}$ equals the maximum size of an antirectangle in $\mathscr{B}$.*

This section comprises the part of the proof of Theorem II that treats some boards by induction on the number of squares. Such boards, which are called *reducible*, have "reducible configurations". A reducible configuration cannot occur in a smallest counterexample to the theorem. Each reducible configuration is accompanied by two constructions. The first produces a smaller "reduced" board $\mathscr{B}'$ from the reducible

board $\mathcal{B}$. The second produces an optimal cover and antirectangle for $\mathcal{B}$ from such a pair for $\mathcal{B}'$. We have two reducible configurations. Each involves a support edge and its associated rectangle. Throughout, assume the relevant support edge is the bottom one.

Tab reduction. Let R be the rectangle associated with a support edge. If the side of R opposite the support edge lies entirely on one edge of the board then $\mathcal{B}$ has a *tab reduction* (Fig. 15a).

$\mathcal{B}'$ is constructed by deleting all the squares in R. In other words, the top and bottom sides of R are collapsed to points and then all pairs of vertical segments that now coincide are deleted. Clearly, $\mathcal{B}'$ is convex. Assume the undeleted squares in $\mathcal{B}$ retain their identity in $\mathcal{B}'$. R itself is collapsed to vertical line l in $\mathcal{B}'$ (FIG. 15b).

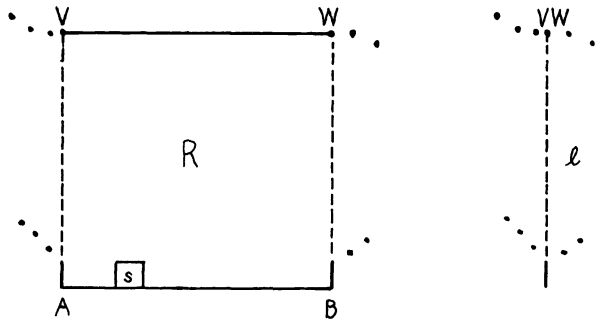


FIG. 15. *Tab reducible configuration before and after reduction.*

From a cover of $\mathcal{B}'$ it is easy to construct a cover of $\mathcal{B}$. Take R and all the rectangles from the former, after stretching horizontally any that cross l. From an antirectangle in $\mathcal{B}'$ construct one in $\mathcal{B}$ by taking all squares in the former and adding any edge square s on the bottom support. The antirectangle in $\mathcal{B}'$ is an antirectangle in $\mathcal{B}$ and remains such when s is added because $N_G(s) = R$ and $\mathcal{B}' = \mathcal{B} \cup R$.

This construction gives an optimal pair in β given an optimal pair for $\mathcal{B}'$; by induction the optimal sizes for $\mathcal{B}'$ are equal, and so they are for $\mathcal{B}$.

Partial tab reduction. For convenience, the conditions for a partial tab reduction are stated for AB the bottom, and GH the right support edge. They can apply to any perpendicular pair of supports.

Conditions for partial tab reduction of $\mathcal{B}$ at AB:

Condition 5.0. $\mathcal{B}$ has no tab reduction.

Condition 5.1. All points in the rectangle associated with GH lie strictly to the right of AB (Fig. 16).

In other words, (5.0) means that the side opposite the support edge of every support edge associated rectangle does not lie entirely on the edge it meets. Condition (5.1) can be restated as $x(B) < x$(left side of rectangle associated with GH) if we assume $x(A) < x(B)$.

Let $R = ABWV$ be the rectangle associated with AB. We examine the consequences of the no tab reduction hypothesis.

First, the top of R cannot be contained in an edge (Fig. 15). Neither can it cover horizontal segments both at its left and right sides (Fig. 17a); otherwise, there would

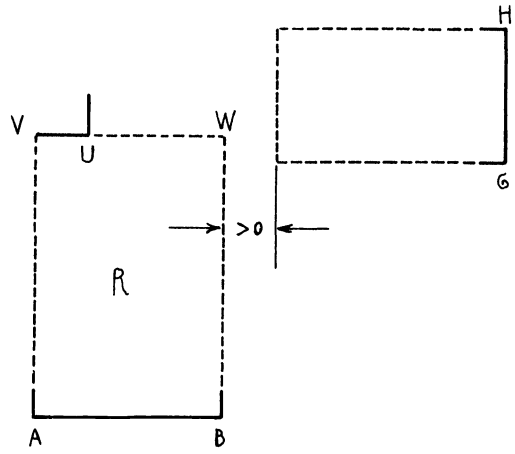


FIG. 16. *Condition for partial tab reduction. No relation between $y(V)$ and $y(G)$ is implied in this figure.*

be a tab reduction at the top support (which must then lie strictly between A and B). We assume without loss of generality that the top of R meets a horizontal edge along UV on the left (Figs. 17b,c). Note that the left support is between A and V (vertically).

Second, we claim that W (upper right corner of R) cannot touch the boundary at all. For if its does, either there is a tab reduction at the top support (Fig. 17b), or BW is part of an edge (Fig. 17c). In the latter case, there is a tab reduction at the left support.

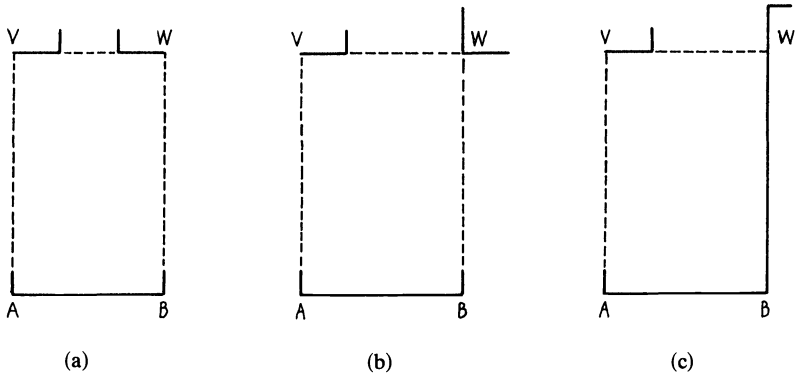


FIG. 17. *Cases eliminated by no tab reduction condition in condition for partial tab reduction.*

We conclude that part of the board looks something like Fig. 18a. The top support does not lie between $x = x(U)$ and $x = x(B)$. BC is the vertical edge at B and AD is the vertical edge at A (D may equal V). Let $y = \min(y(C), y(D))$. $\mathcal{B}'$ is constructed by deleting all squares in $\square A(x(B), y) \cup \square AU$ (Figure 18b). This is equivalent to collapsing all vertical segments with y coordinate between $y(A)$ and y; and all horizontal segments with x coordinate between $x(A) = x(V)$ and $x(U)$. Let $Q = (x(U), y)$.

The construction of a cover for $\mathcal{B}$ from one for $\mathcal{B}'$ is similar to that used for the tab reduction. Take all rectangles from the latter, after stretching any that cross line

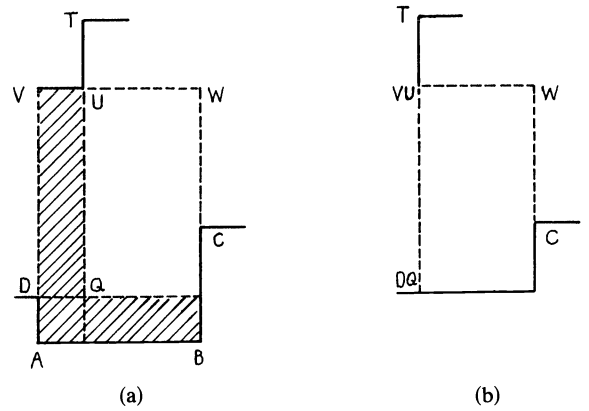


FIG. 18. *Partial tab reducible configuration before and after reduction.*

QU, (which $\square AU$ was collapsed to), and add R. In this cover for $\mathscr{B}$, note $\square QW$ is covered by R and at least one other rectangle. (In Fig. 19 there is a board with no tab reductions and in which Condition 5.1 fails. In it, the analogue of $\square QW$ contains square d which is covered only once in the unique minimum cover. Thus Condition 5.1 is a necessary hypothesis for the above construction to yield an optimal cover.)

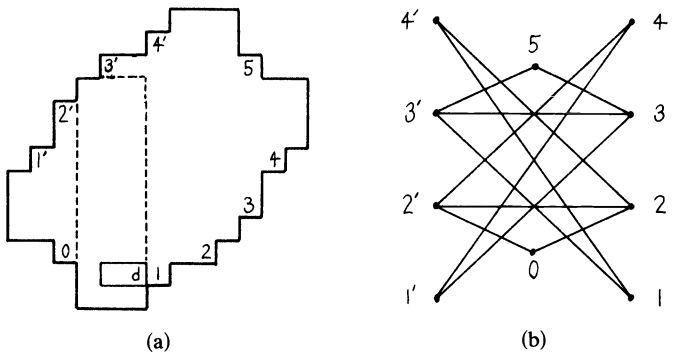


FIG. 19. *Example illustrating the necessity of the partial tab reduction condition (or something like it) for the construction of an optimal cover of β from that of the reduced board.* (a) *has a unique optimal cover consisting of the support edge associated rectangles and* $\square 33'5$, $\square 22'0$, $\square 1'4$ *and* $\square 14'$. *This can be seen from* (b) *which is the induced subgraph of G on the squares not covered by the support edge rectangles. Square d is in the partial tab reduced board but is covered only by the support edge rectangle shown.*

The construction of the antirectangle in $\mathscr{B}$ is more complicated. Let A also denote the corner square at A. Observe $N_G(A) = R$. Let $\mathscr{A}'$ be an antirectangle in $\mathscr{B}'$. If $\mathscr{A}' \cup \{A\}$ is an antirectangle, which means $\mathscr{A}'$ has no square in $\square QW$, we are done. Otherwise $\mathscr{A}' \cap \square QW = \{p\}$ and we claim we can replace p in $\mathscr{A}'$ by some other square to produce a set that remains an antirectangle when A is added to it.

Case 1. $y(D) = y(Q) < y(p) < y(C)$, in other words, Q is below C and $p \in \square QC$. Let f be the leftmost square in $\mathscr{B}$ with $y(f) = y(p)$. The hypothesis and the board structure (convexity) imply that f is left of A and $N_G(f) \subseteq N_G(p)$. Hence $\mathscr{A} = \mathscr{A}' \setminus \{p\} \cup \{A, f\}$ is the desired antirectangle (see Fig. 20a).

Case 2. $y(p) > y(C)$. We prove we can replace p in $\mathscr{A}'$ by at least one of two other squares. Let e be the square $(x(C) + 0.5, y(p))$. Move p right just beyond R.

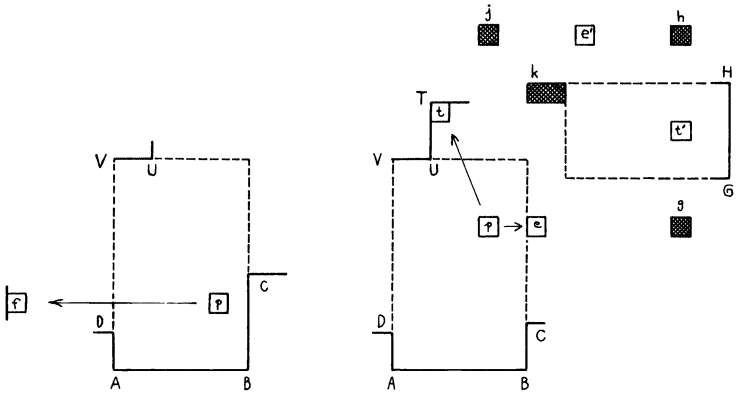


FIG. 20. *Two cases in construction of maximum antirectangles.*

Let t be the corner square at the top end of the vertical edge at U. We show the assumption that neither $\mathscr{A}'\backslash\{p\}\cup\{e\}$ nor $\mathscr{A}'\backslash\{p\}\cup\{t\}$ is an antirectangle leads to a contradiction (see Fig. 20b).

If $\mathscr{A}'\backslash\{p\}\cup\{e\}$ is not an antirectangle, then for some $e'\in\mathscr{A}'$, $(e, e')\in G$ but $(p, e')\notin G$. Since $y(p)=y(e)$, $(x(e'), y(e))\in\beta$ so $j=(x(p), y(e'))\notin\mathscr{B}$. (We use Fact 2.1.)

If $\mathscr{A}'\backslash\{p\}\cup\{t\}$ is not an antirectangle, then for some $t'\in\mathscr{A}'$, $(t, t')\in G$ but $(p, t')\notin G$. The fact that the top support extends right of $x=x(C)$ implies $x(t')>x(C)$. (For if $x(t')<x(C)$ and still $(p, t')\notin G$ and $(t, t')\in G$, then $x(t')<x(p)$ and $(x(p), y(t'))\notin\mathscr{B}$.) Therefore, $(t, t')\in G$ implies $(x(p), y(t'))\in\mathscr{B}$, and so $g=(x(t'), y(p))\notin\mathscr{B}$.

Furthermore, $y(e')>y(t)$ and $y(t')\leqq y(t)$, so $e'\neq t'$. Convexity implies $(x(e'), y(t'))\in\mathscr{B}$, so $\{e', t'\}\subseteq\mathscr{A}'$ implies $h=(x(t'), y(e'))\notin\mathscr{B}$. Some squares in $\mathscr{B}$ (including t') are between g and h, so the rectangle associated with the right support GH must pass between g and h. The condition for partial tab reduction implies that this associated rectangle cannot extend left as far as $x=x(e)$. By convexity, some square $k\notin\mathscr{B}$ has $x(k)=x(e)$, and $y(k)<y(H)<y(h)=y(e')$. This contradicts $(e, e')\in G$.

Hence, for at least one $f\in\{e, t\}$, $\mathscr{A}=\mathscr{A}'\backslash\{p\}\cup\{A, f\}$ is an antirectangle. Suppose we started with an optimal cover and antirectangle for $\mathscr{B}'$. The induction hypothesis (Theorem II) implies that the optimal sizes for $\mathscr{B}'$ are equal, and our construction increases both by one. □

6. Edge covering. In this section, we prove $\alpha(G_{\mathscr{B}}(B))=\theta(G_{\mathscr{B}}(B))$ for convex, irreducible $\mathscr{B}$. More generally, we show $\alpha(G_{\mathscr{B}})=\theta(G_{\mathscr{B}})$ for convex boards that have, at each corner, at most one incident edge that can be partially covered by a maximal rectangle. We assume all rectangles are maximal.

DEFINITION. Suppose CD is an edge, and C is a corner. We say rectangle R *partially covers* CD *at* C if R covers corner square C, but not every edge square on CD.

Our definition means an edge may be partially covered only at an end that is a corner (see Fig. 21).

LEMMA 6.1. *An edge cannot be partially covered at both of its ends.*

Proof. If it were, convexity would imply one of the rectangles is not maximal. □

Now, suppose edge CD is partically covered at C. Let R be the rectangle that partially covers CD at C and that covers as much of CD as possible. Then the side of R perpendicular to CD that does not touch C must cover a segment with an end

point E closest to CD. The edge incident to E that is parallel to CD must reach at least as far as D (see Fig. 21).

Let c' be the edge square on CD closest to C but not in R.

Fact 6.2. Any maximal rectangle that covers c' also covers all of CD (and all of the other edge of C if the other edge cannot be partially covered at C).

Fact 6.3. $N_G(C) \supseteq N_G(c')$.

We call c' a "proxy" for C.

LEMMA 6.4. *Let $\mathcal{B}$ be a convex board for which at each corner there is at most one incident edge that can be partially covered at that corner. Then* $\alpha(G_{\mathcal{E}}(\mathcal{B})) = \theta(G_{\mathcal{E}}(\mathcal{B}))$.

Proof. Let $\mathcal{P} \subseteq \mathcal{B}$ consist of the unique proxy for each corner square at which an edge can be partially covered, plus all corner squares at which neither incident edge can be partially covered. Fact 6.3 implies that $G^* = G_{\mathcal{P}}(\mathcal{B})$ is isomorphic to an arc deleted subgraph of $G_{\mathcal{E}}$. We claim this subgraph of $G_{\mathcal{E}}$ satisfies the hypothesis of Lemma 4.2. Consider any 3 clique $T = \{a, b, c\}$ in $G_{\mathcal{E}}$. If $b \in T$ is the square at the right angle of a triangle of squares, that is, the degree 2 square of Lemma 4.2, it cannot be a corner square that is replaced by a proxy. On the other hand, let $c \in T$ be a corner square that is replaced by a proxy c'; then neither (a, c') nor (b, c') is an arc in G^*, or else both are. Therefore, Lemma 4.2 implies $\theta(G^*) = \alpha(G^*)$. Fact 6.2 implies any set of maximal rectangles corresponding (using Lemma 1.1) to a clique cover of G^* covers all the edges. □

LEMMA 6.5. *If β is convex and irreducible,* $\alpha(G_{\mathcal{E}}(\mathcal{B})) = \theta(G_{\mathcal{E}}(\mathcal{B}))$.

Proof. Suppose the hypothesis for Lemma 6.4 were false for $\mathcal{B}$. If $\mathcal{B}$ were as in Fig. 22, there would be a tab reduction between U and V. Hence $\mathcal{B}$ must be as in Fig. 23.

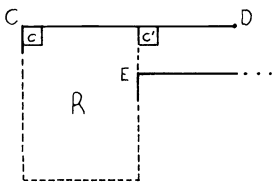


FIG. 21

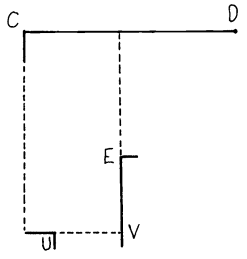


FIG. 22

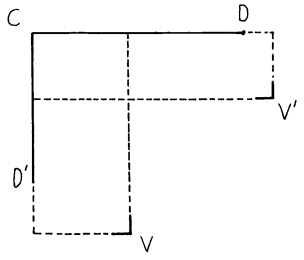


FIG. 23

Since there are no tab reductions in this $\mathcal{B}$, (by convexity) some of the right support edge must be above $y = y(D)$. Hence, the left side of the right support edge associated rectangle R cannot be left of $x = x(D)$. Let B be the right end of the

bottom support. $x(B)<x(V)$ (again by convexity) and $x(V)<x(D)$, so B is strictly left of the left side of R. This is Condition 5.1 for partial tab reduction at the bottom support, and so by contradiction, the hypothesis of Lemma 6.4 is true. □

7. Whole covering of irreducible boards. This section concludes the proof of Theorem II. We prove that if every optimal edge cover of an irreducible board $\mathcal{B}$ fails to cover every square, then an optimal whole cover can be obtained by adding one rectangle to an optimal edge cover. Suppose $\mathcal{B}$ is such a board. Consider an optimal edge cover **C** (by maximal rectangles) that covers the maximum number of squares. Let z be an uncovered square.

There are three steps. The first is to establish the structure of $\mathcal{B}$. The second is to show the squares not covered by **C** can all be covered by one rectangle. The third is to prove optimality by constructing an antirectangle that contains z and one square for each rectangle of **C**.

Step 1. *Structure of* $\mathcal{B}$. In each of the four directions, there is an edge square on the same line as z. Let these squares be r_i, r_i', $i = 1, 2$ as shown in Figs. 24–25. Consider

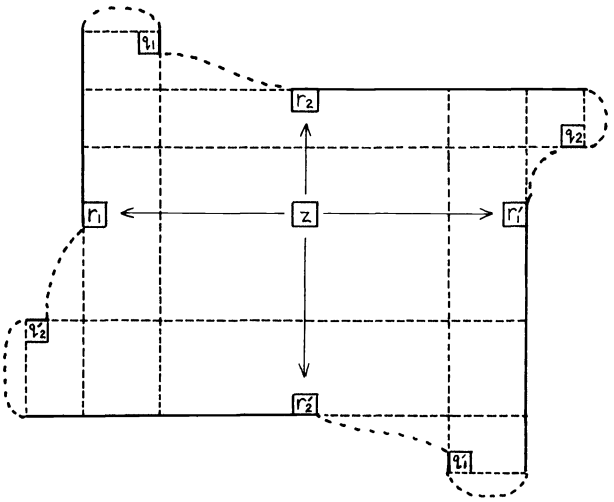


FIG. 24. *Dotted lines indicate polygon boundary schematically. In this case, the board is reducible.*

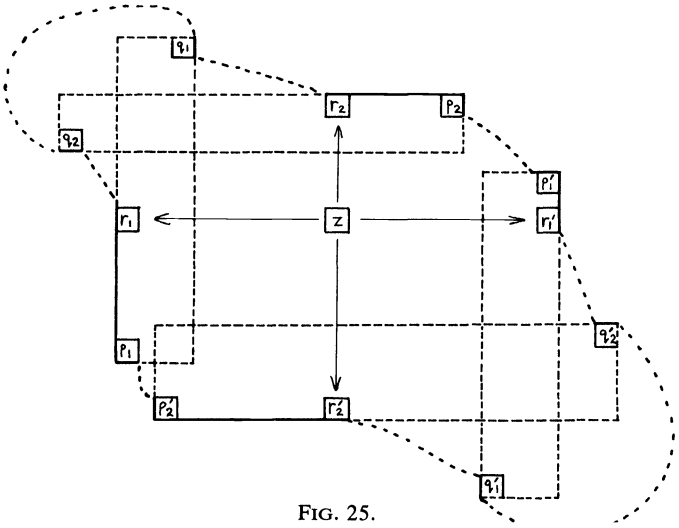


FIG. 25.

a pair of rectangles in **C** that cover r_1 and r_1'. We can assume without loss of generality that the "inner" (right) side of the (left) rectangle covering r_1 meets a vertical segment q_1 at its upper end, and the "inner" (left) side of the (right) rectangle covering r_i' meets a segment q_1' at its lower end. Each inner side must meet a segment because the rectangles are maximal. The two parallel inner sides could not each meet a segment at the same end because this would violate convexity.

Apply the above argument to the rectangles that cover r_2 and r_2'. Now there are two possibilities. The first, that the "inner" (bottom) side of the (top) rectangle which covers r_2 meets a segment on the right, is shown in Fig. 24. Here, convexity implies the solidly drawn parts of the rectangle sides are parts of edges, so there is a tab reduction at each support edge; this contradicts the assumption that $\mathcal{B}$ is irreducible. Hence our four rectangles must meet the boundary of $\mathcal{B}$ as in Fig. 25. Taking into account the general structure of convex boards, we conclude:

Let $\mathbf{K}_i(\mathbf{K}_i')$ be the set of rectangles in **C** that contain squares that lie between z and $r_i(r_i')$. Each rectangle in $\mathbf{K}_i$ ($\mathbf{K}_i'$) covers two segments q_i and $p_i(q_i'$ and $p_i')$ as the rectangles shown in Figs. 25 and 28 do. Here, segments labeled q block extension of the rectangle toward z and segments labeled p block extension away from q.

For example, any rectangle in $\mathbf{K}_1$ covers a horizontal segment like p_1 at its lower left corner and a vertical segment like q_1 at its upper right corner.

The boundary segments and vertices are cyclically ordered counterclockwise (CCW). When e and f are two edge squares $[e, f]$ denotes all the edge squares on any segment on the CCW path from a segment on e to a segment of f. $[e, f) = [e, f]\backslash\{e\}$, etc.

Step 2. *Covering uncovered squares with one more rectangle.* Consider the four support edge rectangles. The result of Step 1 and the irreducibility hypothesis imply these rectangles must be as in Fig. 26. These rectangles "cross" at D and B and intersect at A and C, as shown. By convexity, $\square ABCD \subseteq \mathcal{B}$. We claim all squares not covered by **C** lie in $\square ABCD$. To prove this, we note that the nonsupport (maximal) rectangles in edge cover **C** also cover the region between the boundary and the support rectangles. For example, again looking at Fig. 26, we see that any rectangle that covers any corner square f_1 or f_2 in $[E_1, E_2]$ or $[E_4, E_5]$ covers $\square f_1 D$ or $\square f_2 A$ respectively.

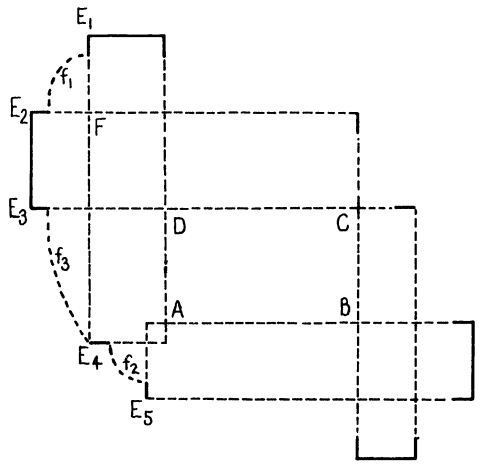


FIG. 26. *The support edge rectangles illustrating Step* 2. *The rectangle may overlap (as A) or touch (as C) at A and C. All uncovered squares must be in* $\square ABCD$.

A rectangle covering edge square f_3 in $[E_3, E_4]$ covers all the squares on the horizontal line from f_3 to DA.

Step 3. *Constructing maximum antirectangles.*

DEFINITION. Let $\mathbf{C}$ be a set of rectangles. A square $s \in \mathcal{B}$ is *critically covered* when it is in only one rectangle in $\mathbf{C}$. Two squares are *matched* by $R \in \mathbf{C}$ if both are critically covered by R.

We note that $\mathbf{C}$ is a minimum edge cover, so every rectangle in $\mathbf{C}$ critically covers at least one edge square. If $\mathcal{A}$ is an antirectangle and $|\mathcal{A}| = |\mathbf{C}|$, $\mathcal{A}$ consists only of critically covered squares.

Let us choose one rectangle from each $\mathbf{K}_i$, $\mathbf{K}_i'$ that is closest to square z from among all rectangles in $\mathbf{K}_i$, $\mathbf{K}_i'$. These four rectangles and the two boundary segments p, q on each we described in Step 1 and are displayed in Fig. 28. Let $\mathbf{K} = \mathbf{K}_1 \cup \mathbf{K}_2 \cup \mathbf{K}_1'\mathbf{K}_2'$. We divide the proof into numbered assertions.

1) If an edge square $e \in [s_1, s_2] \cup [s_1', s_2']$ on edge E is critically covered, so is the corner square c on E. For the maximal rectangle covering c is unique and contains e. Let S be the set of corner squares in $[s_1, s_2] \cup [s_1', s_2']$.

2) Let $\bar{z} = \{s | s$ is an edge square and $(z, s) \notin G\}$. We claim that every rectangle in $\mathbf{K}$ critically covers a square in $\bar{z}$. Every rectangle R in, for example, $\mathbf{K}_1$ critically covers some edge squares. If R critically covers some squares in (r_2, r_1) we are done. Otherwise, R must critically cover some edge squares $E \subseteq [r_1, r_2')$ and no edge squares elsewhere. Now, if R did not critically cover any interior squares above line r_1z, we could replace R in $\mathbf{C}$ by a maximal rectangle R_1 obtained as follows. Shrink the top of R down to r_1z, then extend the right side as far as possible, and then finally extend the top to make the rectangle maximal. The result is a minimum edge cover that covers more squares than $\mathbf{C}$. On the other hand, suppose R critically covered a set of interior squares H above line r_1z. The same replacement process for R still produces a minimum edge cover that covers z. If this new edge covers H, again the maximality of $\mathbf{C}$ is contradicted. If not, the uncovered squares in H are squares not covered by a minimum edge cover of maximal rectangles and so they lie (along with z) in $\square ABCD$ as we showed in Step 2. Therefore, there exists a rectangle R_2 covering $E \cup H \cup \{z\}$ (see Fig. 27).

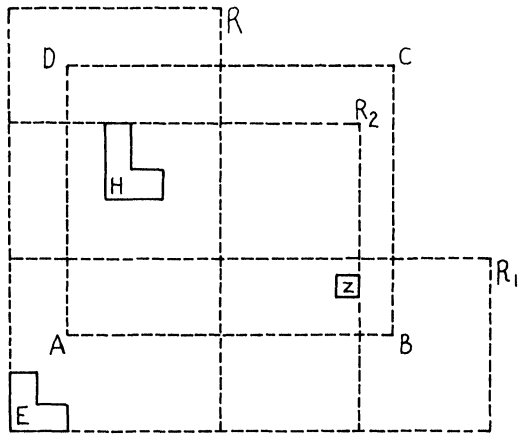


FIG. 27. *The maximality of the set covered by* $\mathbf{C}$ *implies every* $R \in K$ *critically covers a square in z. If not, replace R by* R_1 *or* R_2 *(see Step* 2).

3) Define the sets of edge squares shown in Fig. 28,

$$R_1 = (r_2, q_1), \qquad R_2 = (q_2, r_1),$$
$$Q_1 = [q_1, s_1) \qquad Q_2 = (s_2, q_2].$$

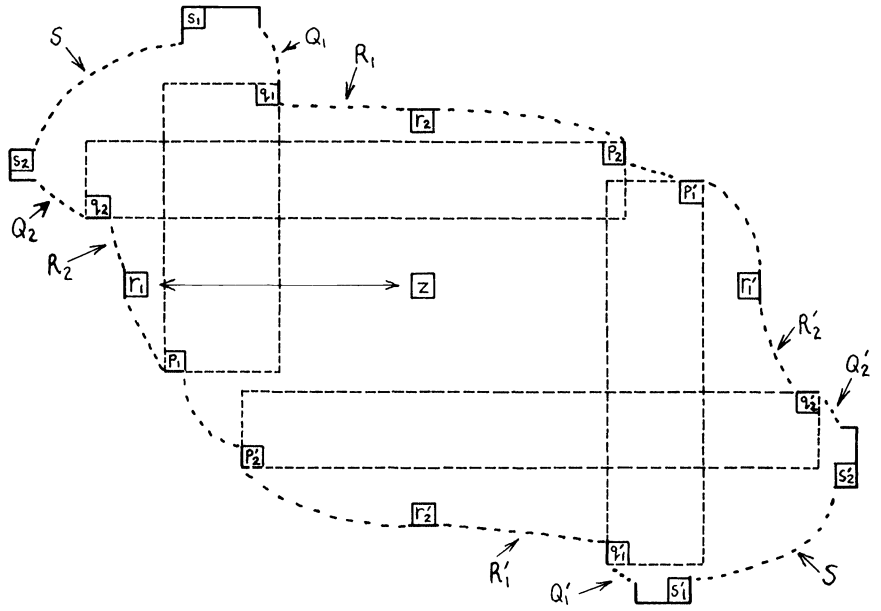


FIG. 28.

We claim that every rectangle in $\mathbf{K}_i$ critically covers a square in $S \cup Q_i$. By 2), $R \in \mathbf{K}_i$ critically covers a square in $S \cup Q_i \cup R_i$. If it covered a square in R_i, however, this would violate the assumption that $\square p_1 q_1$ is a rectangle in $\mathbf{K}_i$ closest to z.

4). If $R \in \mathbf{C}$ covers a square in R_i, then it must critically cover a square in $R_{i'} \cup Q_{i'} \cup S$, where $i' = 3 - i$. Such $R \notin \mathbf{K}$. Otherwise, we can give a proof similar to that in 2) to contradict the maximality of $\mathbf{C}$ or the choice of $\square p_i q_i$.

We conclude from 1)–4) that every rectangle $R \in \mathbf{C}$ satisfies exactly one of two possibilities and define $a : \mathbf{C} \to \mathcal{B}$ thereby:

i. R critically covers a square in $S \cup Q_1 \cup Q_2$ or $S \cup Q_1' \cup Q_2'$. If it critically covers some $s \in S$, set $a(R) = s$, otherwise set $a(R)$ to any square in $\cup Q$ critically covered by R.

ii. If i does not hold, then R matches a square $a = a(R) \in R_1$ with a square in R_2, or $a = a(R) \in R_1'$ with a square in R_2'. This follows from 4. (Note the asymmetry. $a(R)$ is always chosen from R_1 or R_1'.)

Let $\mathcal{A} = \{a(R) | R \in \mathbf{C}\}$. $a(R) \in \mathcal{A}$ is always critically covered by R so $|\mathcal{A}| = |\mathbf{C}|$. As defined, $\mathcal{A} \subseteq \bar{z}$, so as long as $\mathcal{A}$ is an antirectangle $\mathcal{A} \cup \{z\}$ is the desired maximum antirectangle. Clearly, S, $Q_i \cup R_i$, and $Q_i' \cup R_i'$ are each antirectangles (consider corner types). We conclude the proof that $\mathcal{A}$ is an antirectangle by showing each of the three remaining possible ways for two squares in a common rectangle to be both in $\mathcal{A}$ leads to contradiction.

5) Suppose $\{s, q\} \subseteq \mathcal{A}$ with $s \in S$, $q \in (\cup Q) \cup (\cup R)$, and $(s, q) \in G$. This cannot happen because (by 1) s is a corner square and is contained in a unique maximal rectangle, so q cannot be critically covered.

6) Suppose say, $q \in Q_1, q' \in Q_2$ and $(q, q') \in G$. The board structure implies $(q_1, q_2) \in G$. $\mathbf{C} \backslash \{\square p_1 q_1, \square p_2 q_2\} \cup \{\square p_1 p_2, \square q_1 q_2\}$ covers more than $\mathbf{C}$, including z. This contradicts the maximality of $\mathbf{C}$.

7) Suppose say, $\{r, q\} \subseteq \mathscr{A}$ with $r \in R_1$, $q \in Q_2$, and $(r, q) \in G$. Then $(r, q_2) \in G$. $r \in \mathscr{A}$ implies $r = a(R)$ for some $R = \square rr'$ that satisfies ii; that is, R matches $r' \in R_2$ with r. We do another switch, $\mathbf{C} \backslash \{\square rr', \square q_2 p_2\} \cup \{\square r' p_2, \square r q_2\}$. Again, the maximality of the set covered by $\mathbf{C}$ is contradicted.

REFERENCES

[1] C. BERGE, *Perfect graphs*, in Studies in Graph Theory Part 1, D. R. Fulkerson, ed., Studies in Mathematics 11, Mathematical Association of America, 1975, pp. 1–22.

[2] R. BRUALDI, *Transversal theory and graphs*, in Studies in Graph Theory, Part 1, D. R. Fulkerson, ed., Studies in Mathematics 11, Mathematical Association of America, 1975, pp. 23–88.

[3] F. R. K. CHUNG, Personal communication, February, 1979.

[4] C. C. LIU, *Introduction to Combinatorial Mathematics*, McGraw-Hill, New York, 1968.

[5] W. MASEK, Personal communication, June, 1979.

[6] D. R. WOODALL, *Minimax theorems in graph theory*, in Selected Topics in Graph Theory, L. W. Beineke and R. J. Wilson, eds., Academic Press, New York, 1978, pp. 237–269.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 4, December 1981

0196-5212/81/0204-0006 $01.00/0

MINIMEAN LOCATION OF DIFFERENT FACILITIES ON A LINE NETWORK*

B. L. HULME† AND P. J. SLATER†

Abstract. The m-mean median problem for a network N is introduced in the context of locating m different facilities on m of the vertices in $V(N)$. If we let $d(v, w)$ denote the distance in N between vertices v and w, the problem is to select an m-set $S \subseteq V(N)$ with complement $\bar{S} = V(N) - S$ so as to minimize the sum $\sum_{u \in S} \sum_{v \in \bar{S}} d(u, v)$. That is, one seeks to partition $V(N)$ into two sets, a set S of "facility vertices" and a set $\bar{S}$ of "customer vertices," so as to minimize the *average* distance between a facility and a customer. Complete results are given for the special case of a line network.

1. Introduction. The center and median problems introduced by Hakimi [3] and, more generally, the m-center and m-median problems of optimal facility location in a network have been extensively studied. (See, for example, [1], [2], [4]–[8].) In these problems the customers and their locations are perceived as given and fixed, and sites for new facilities are to be selected with respect to these known customer locations under various "optimality" criteria.

The architect for a factory or a city planner, however, may be able to preplan both the service facility and customer locations. Here one problem of this nature is introduced for arbitrary networks, and the solution is presented for line networks. Specifically, assume there is a network N whose vertex set $V(N)$ will be partitioned into a set of size m for m different facilities and a set of $|V(N)| - m$ customer locations. Upon each service facility an equal demand is placed by each customer, and each customer places an equal service requirement upon each of the m facilities. We seek to minimize the average distance between a customer location and a service location, which is equivalent to minimizing the sum of the $m(|V(N)| - m)$ distances of a customer to a facility.

Let N be a simple connected network with vertex set $V(N) = \{v_0, v_1, \cdots, v_n\}$ and edge set $E(N)$. If $e = (v_i, v_j)$ is the edge connecting v_i and v_j, then $L(e)$ or $L(v_i, v_j)$ will denote the (positive) length of e. A finite nonnull sequence $P = u_0 e_1 u_1 e_2 u_2 \cdots e_k u_k$ whose distinct terms are alternately vertices and edges with $e_i = (u_{i-1}, u_i)$ is called a $u_0 - u_k$ *path* of length $L(P) = \sum_{i=1}^{k} L(e_k)$. The distance between vertices v_i and v_j, denoted $d(v_i, v_j)$, equals the smallest length of a $v_i - v_j$ path. For each $S \subseteq V(N)$, let $\bar{S} = V(N) - S$, and let

$$M(S) = \sum_{v_i \in S} \sum_{v_j \in \bar{S}} d(v_i, v_j). \tag{1}$$

The m-mean median problem. For $1 \leqq m \leqq n$, let

$$M_m(N) = M_m(V(N)) = \min \{M(S): S \subseteq V(N), |S| = m\}. \tag{2}$$

Find an m-set $S \subseteq V(N)$ for which $M(S) = M_m(N)$, where such a set S is referred to as an *m-mean median.*

For example, if S is the set of darkened vertices in Fig. 1, then, for each edge having length one, each value of $\sum_{v_j \in \bar{S}} d(v_i, v_j)$ is as indicated for each $v_i \in S$, and $M_6(T) = M(S) = 102$.

* Received by the editors May 27, 1980, and in revised form March 10, 1981.

† Applied Mathematics Department, Sandia National Laboratories, Albuquerque, New Mexico 87185.
This article was sponsored by the U.S. Department of Energy under contract DE-AC04-76DP00789.

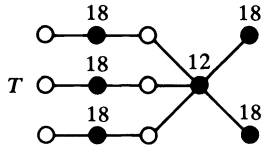


FIG. 1. *Tree T with* $M_6(T) = 102$.

Observation 1. S is an m-mean median if and only if $\bar{S}$ is an $(n+1-m)$-mean median.

In [9], [10] competitive location theory problems were introduced for networks, and the notation $V(u, v) = \{w \in V(N) : d(w, u) < d(w, v)\} - \{u\}$ was used. That is, $V(u, v)$ is the set of vertices in $V(N)$, other than u itself, which are closer to u than to v. Note that if T is a tree network containing edge (u, v) then $T-(u, v)$ has two components with vertex sets $V(u, v) \cup \{u\}$ and $V(v, u) \cup \{v\}$. The fundamental result to be used extensively in § 2 is the next lemma.

LEMMA 2. *If S is an m-set from vertex set $V(T)$ of a tree T, $(u, v) \in E(T)$, $u \in S$, $v \in \bar{S}$, $C(u) = |S \cap V(u, v)|$, $\bar{C}(u) = |\bar{S} \cap V(u, v)|$, $C(v) = |S \cap V(v, u)|$ and $\bar{C}(v) = |\bar{S} \cap V(v, u)|$, then $M(S-u+v) = M(S) + L(u, v)(\bar{C}(u) + C(v) - \bar{C}(v) - C(u))$.*

Proof. Let $R = S - u$ and $Q = \bar{S} - v$. Then

$$M(S) = d(u, v) + \sum_{r \in R} \sum_{t \in Q} d(r, t) + \sum_{r \in R} d(r, v) + \sum_{t \in Q} d(u, t)$$

and

$$M(S-u+v) = d(u, v) + \sum_{r \in R} \sum_{t \in Q} d(r, t) + \sum_{r \in R} d(r, u) + \sum_{t \in Q} d(v, t).$$

Thus

$$\begin{aligned} M(S-u+v) - M(S) &= \sum_{r \in R} (d(r, u) - d(r, v)) + \sum_{t \in Q} (d(v, t) - d(u, t)) \\ &= \sum_{r \in R \cap V(u, v)} (d(r, u) - d(r, v)) + \sum_{r \in R \cap V(v, u)} (d(r, u) - d(r, v)) \\ &\quad + \sum_{t \in Q \cap V(u, v)} (d(v, t) - d(u, t)) + \sum_{t \in Q \cap V(v, u)} (d(v, t) - d(u, t)) \\ &= L(u, v)(-C(u) + C(v) + \bar{C}(u) - \bar{C}(v)). \end{aligned}$$

□

COROLLARY 3. *For u, v, $C(u)$, $\bar{C}(u)$, $C(v)$ and $\bar{C}(v)$ as in Lemma 2, if S is an m-mean median of tree T, then $\bar{C}(u) + C(v) \geqq \bar{C}(v) + C(u)$.*

2. The m-mean medians of line networks. In this section we restrict our attention to line networks P_n with $V(P_n) = \{v_0, v_1, v_2, \cdots, v_{n-1}\}$ and $E(P_n) = \{(v_{i-1}, v_i) : 1 \leqq i \leqq n-1\}$. First to be considered will be line networks with an even number of vertices, say $n = 2p$. By Observation 1, it can be assumed that $m \leqq p$.

THEOREM 4. *If S is an m-mean median of P_{2p} and $m < p$, then v_0 and v_{2p-1} must be in $\bar{S}$. Furthermore, S is an m-mean median of P_{2p} if and only if S is an m-mean median of $P' = v_1, v_2, \cdots, v_{2p-2}$.*

Proof. Assume S is an m-mean median with $v_0 \in S$, and let t be the smallest value for which $v_t \in \bar{S}$. With $u = v_{t-1}$ and $v = v_t$ and the notation of Lemma 2, $C(u) = t-1$,

$\bar{C}(u)=0$, $C(v)=m-t$ and $\bar{C}(v)=2p-m-1$. Thus we have, by Corollary 3,

$$0+(m-t)\geqq(2p-m-1)+(t-1),$$

$$2m\geqq 2p+2t-2$$

and

$$2m\geqq 2p.$$

Since this contradicts the fact that $m<p$, we have $v_0\in\bar{S}$. Similarly, v_{2p-1} must be in $\bar{S}$.

Let $S\subseteq V(P')$ be any m-set, and let $\bar{S}'=V(P')-S=\bar{S}-\{v_0, v_{2p-1}\}$. If $v_i\in S$, then $\sum_{w\in\bar{S}}d(v_i, w)=D+\sum_{w\in\bar{S}'}d(v_i, w)$, where $D=d(v_0, v_{2p-1})=\sum_{e\in E(P_{2n})}L(e)$. This implies that $M(S)$ in P_{2p} equals $M(S)$ in P' plus $m\cdot D$. Consequently S is an m-mean median in P' if and only if it is an m-mean median in P_{2p}. □

THEOREM 5. *If $m=p$, then S is an m-mean median of P_{2p} if and only if for $0\leqq s\leqq m-1$ each 2-set $\{v_{2s}, v_{2s+1}\}$ contains one element in S and one element in $\bar{S}$.*

Proof. Suppose S is an m-mean median of P_{2p}, and hence so is $\bar{S}$. Assume v_0 and v_1 are both in one of these m-mean medians, say $\{v_0, v_1\}\subseteq S$. As in the proof of Theorem 4, letting t be the smallest value for which $v_t\in\bar{S}$, by Corollary 3 one obtains the inequality $2m\geqq 2(p-1+t)$. Now, however, $t\geqq 2$ implies $m\geqq p+1$ which contradicts the fact that $m=p$. Thus $|S\cap\{v_0, v_1\}|=1=|\bar{S}\cap\{v_0, v_1\}|$. A simple induction on s will show that each $\{v_{2s}, v_{2s+1}\}$ contains one element in S and one element in $\bar{S}$.

To complete a proof of the theorem it must be shown that any S with $|S\cap\{v_{2s}, v_{2s+1}\}|=1$ for $0\leqq s\leqq m-1$ is an m-mean median, or, equivalently, that all 2^m such m-sets S have the same value of $M(S)$. Let S_1 be such a set and assume that $v_{2r}\in S_1$ for some $0\leqq r\leqq m-1$. It will suffice to show that $S_2=S_1-v_{2r}+v_{2r+1}$ satisfies $M(S_1)=M(S_2)$. Applying Lemma 2 with $u=v_{2r}$ and $v=v_{2r+1}$, we have $C(v_{2r})=r=\bar{C}(v_{2r})$ and $C(v_{2r+1})=p-r-1=\bar{C}(v_{2r+1})$, and so $M(S_2)=M(S_1)+L(v_{2r}, v_{2r+1})\times(p-1-(p-1))=M(S_1)$. □

In contrast to the multisolutions for P_{2p}, the m-mean median of P_{2p+1} is unique. Since arguments similar to those used to prove Theorems 4 and 5 would suffice to prove the next theorem, its proof is omitted.

THEOREM 6. *Let $Q_1=(v_p, v_{p-2}, v_{p+2}, v_{p-4}, v_{p+4}, \cdots)$ and $Q_2=(v_{p-1}, v_{p+1}, v_{p-3}, v_{p+3}, \cdots)$. If $1\leqq m\leqq p$ then the m-mean median of P_{2p+1} is unique and consists of*

1. *the first m vertices of Q_1 if m is odd, or*
2. *the first m vertices of Q_2 if m is even.*

Examples of the 4-mean medians of some line networks are presented in Fig. 2. Recall that the 4-mean median of a line network with an even number of vertices is not unique, and note that we have shown that the determination of the m-mean median is independent of the edge lengths in the line network.

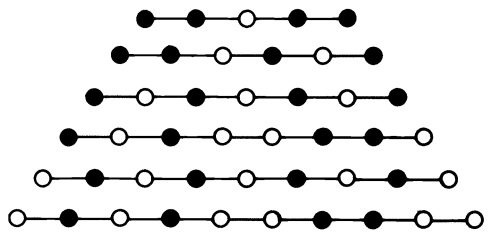


FIG. 2. *4-mean medians of the line networks P_5, P_6, P_7, P_8, P_9 and P_{10}.*

3. Computing the distance sum. In this section a recurrence relation is derived for $M_m(P_n)$, the sum of the distances from an m-mean median to the other vertices in an n-vertex line network with unit edge lengths. (Let P_n denote the line network with vertices $v_0, v_1, \cdots, v_{n-1}$ and edges $e_i = (v_{i-1}, v_i)$ with $L(e_i) = 1$ for $1 \leqq i \leqq n-1$.) The solution of the relation yields a formula for this distance sum in terms of m and n.

The darkened vertices in Fig. 3 show m-mean medians for P_{2p-1} and P_{2p}. The first follows from Theorem 6 with p replaced by $p-1$, since for m odd the first m vertices of Q_1 are $\{v_{p-1}, v_{p-3}, v_{p+1}, v_{p-5}, v_{p+3}, \cdots, v_{p-m}, v_{p+m-2}\}$, and for m even the first m vertices of Q_2 are $\{v_{p-2}, v_p, v_{p-4}, v_{p+2}, \cdots, v_{p-m}, v_{p+m-2}\}$. The m-mean median for P_{2p} results from the fact that by Theorem 4 the pairs $\{v_0, v_{2p-1}\}, \{v_1, v_{2p-2}\}, \cdots, \{v_{p-m-1}, v_{p+m}\}$ do not belong to the m-mean median and by Theorem 5 we may choose every other remaining vertex beginning with v_{p-m}, i.e., $\{v_{p-m}, v_{p-m+2}, \cdots, v_{p+m-2}\}$, to be the m-mean median.

Since $S = \{v_{p-m}, v_{p-m+2}, \cdots, v_{p+m-2}\}$ is an m-mean median for both P_{2p-1} and P_{2p}, and since P_{2p} differs from P_{2p-1} only by having one additional vertex, v_{2p-1}, we have

$$M_m(P_{2p}) - M_m(P_{2p-1}) = \sum_{v_j \in S} d(v_{2p-1}, v_j) = mp. \tag{3}$$

P_{2p} in Fig. 3 has an alternate m-mean median obtained by shifting the darkened vertices one place to the right (Theorem 5). Fig. 4 shows that P_{2p+1} and P_{2p} share this new m-mean median S'. The same argument as before leads to

$$M_m(P_{2p+1}) - M_m(P_{2p}) = \sum_{v_j \in S'} d(v_{2p}, v_j) = mp. \tag{4}$$

Equations (3) and (4) may be summarized as

$$M_m(P_n) - M_m(P_{n-1}) = m\left\lfloor \frac{n}{2} \right\rfloor, \qquad n \geqq 2m. \tag{5}$$

Only $n \geqq 2m$ is considered here because $M_k(P_n) = M_{n-k}(P_n)$, $1 \leqq k \leqq n-1$.

For fixed m, (5) is a linear first-order difference equation in n. Accordingly $M_m(P_n)$ must be the sum of a particular solution of (5) and a homogeneous solution h which depends only on m. A particular solution will be m times a quadratic in $\lfloor n/2 \rfloor$, and inspection shows that $m \lfloor n/2 \rfloor \lfloor (n+1)/2 \rfloor$ is such a solution. Hence

$$M_m(P_n) = m\left\lfloor \frac{n}{2} \right\rfloor \left\lfloor \frac{n+1}{2} \right\rfloor + h(m). \tag{6}$$

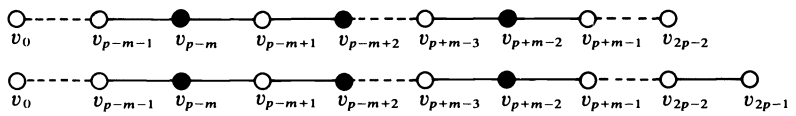


FIG. 3. *m-mean medians of P_{2p-1} and P_{2p}.*

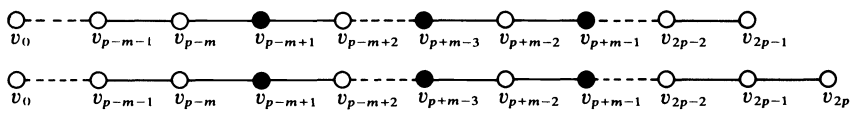


FIG. 4. *m-mean medians of P_{2p} and P_{2p+1}.*

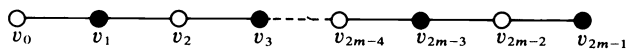


FIG. 5. *m-mean median for* P_{2m}.

In order to determine $h(m)$, an initial condition for $M_m(P_{2m})$ is needed. P_{2m} is shown in Fig. 5. Summing the distances from each open vertex to each darkened vertex starting with (v_0, v_1) yields

$$\begin{aligned} M_m(P_{2m}) &= [1+3+\cdots+(2m-1)]+[1+1+3+\cdots+(2m-3)] \\ &\quad +[3+1+1+3+\cdots+(2m-5)]+\cdots \\ &\quad +[(2m-3)+(2m-5)+\cdots+3+1+1] \\ &= \frac{m(2m^2+1)}{3}. \end{aligned} \tag{7}$$

Therefore, (6) and (7) imply that for $n = 2m$

$$m^3 + h(m) = \frac{m(2m^2+1)}{3},$$

so that

$$h(m) = \frac{m(1-m^2)}{3}.$$

Consequently, the solution of (5) and (7) is given by

$$M_m(P_n) = m\left(3\left\lfloor\frac{n}{2}\right\rfloor\left\lfloor\frac{n+1}{2}\right\rfloor - m^2 + 1\right)\Big/3, \qquad 1 \leqq m \leqq \frac{n}{2}. \tag{8}$$

The other values are obtained from (8) and

$$M_m(P_n) = M_{n-m}(P_n), \qquad \frac{n}{2} < m \leqq n-1. \tag{9}$$

REFERENCES

[1] P. M. DEARING, R. L. FRANCIS AND T. J. LOWE, *Convex location problems on tree networks*, Oper. Res., 24 (1976), pp. 628–642.
[2] A. J. GOLDMAN, *Minimax location of a facility in a network*, Transportation Sci., 6 (1972), pp. 407–418.
[3] S. L. HAKIMI, *Optimum locations of switching centers and the absolute centers and medians of a graph*, Oper. Res., 12 (1964) pp. 450–459.
[4] S. L. HAKIMI AND S. N. MAHESHUARI, *Optimum locations of centers in networks*, Oper. Res. 20 (1972), pp. 967–973.
[5] S. L. HAKIMI, E. F. SCHMEICHEL AND J. G. PIERCE, *On p-centers in networks*, Transportation Sci., 12 (1978), pp. 1–15.
[6] O. KARIV AND S. L. HAKIMI, *An algorithmic approach to network location problems* II: *The p-medians*, SIAM J. Appl. Math., 37 (1979), pp. 539–560.
[7] E. MINIEKA, *The m-center problem*, SIAM Rev. 12 (1970), pp. 138–139.
[8] ———, *The centers and medians of a graph*, Oper. Res., 25 (1977), pp. 641–650.
[9] P. J. SLATER, *Maximin facility location*, J. of Research of the N.B.S., 79B (1975), pp. 107–115.
[10] ———, *Central vertices in a graph*, in Proc. 7th S.E. Conf. on Combinatorics, Graph Theory and Computing, F. Hoffman, et al., eds., Utilitas Mathematica Publishing, Inc., Winnipeg 1976, pp. 487–497.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 4, December 1981

0196-5212/81/0207-0007 $01.00/0

INTEGER ROUNDING FOR POLYMATROID AND BRANCHING OPTIMIZATION PROBLEMS*

S. BAUM† AND L. E. TROTTER, JR.‡

Abstract. Where matrix $M \geqq 0$ and vector $w \geqq 0$ have rational entries, define $r^*(w) = \max\{1 \cdot y : yM \leqq w, y \geqq 0\}$, $z^*(w) = \max\{1 \cdot y : yM \leqq w, y \geqq 0, y \text{ integral}\}$. *Integer round-down* holds for M if, for all integral $w \geqq 0$, $\lfloor r^*(w) \rfloor = z^*(w)$. Similarly, when $\lceil r_*(w) \rceil = z_*(w)$ for all integral $w \geqq 0$, where $r_*(w) = \min\{1 \cdot y : yM \geqq w, y \geqq 0\}$, $z_*(w) = \min\{1 \cdot y : yM \geqq w, y \geqq 0, y \text{ integral}\}$, *integer round-up* holds for M. The integer round-down and round-up properties are shown to hold for certain matrices related to integral polymatroids and branchings in directed graphs.

1. Introduction. Let M be a nonvacuous $m \times n$ matrix of nonnegative rationals and consider the following linear and related integer programming problems parameterized by the nonnegative rational n-vector w:

$$\mathrm{P}(w) \qquad \{\max 1 \cdot y : yM \leqq w, y \geqq 0\}$$

$$\mathrm{P_I}(w) \qquad \{\max 1 \cdot y : yM \leqq w, y \geqq 0, y \text{ integer}\}.$$

Here 1 and 0 are appropriately dimensioned vectors of ones and zeros, respectively, and $1 \cdot y = \sum_{i=1}^{m} y_i$. It is clear that $\mathrm{P}(w)$ and $\mathrm{P_I}(w)$ are feasible and that these programs have bounded objective value if and only if M has no row consisting entirely of zeros. We will assume that M has no zero rows and say that the *integer round-down* (IRD) *property* holds for M if, for each nonnegative integral n-vector w, the optimal objective value of the program $\mathrm{P_I}(w)$ is given by rounding the optimal objective value for $\mathrm{P}(w)$ down to the nearest integer. Similarly, the *integer round-up* (IRU) *property* holds for M if, for each nonnegative integral n-vector w, the optimal value of $\mathrm{C_I}(w)$ is obtained by rounding the optimal value for $\mathrm{C}(w)$ up to the nearest integer, where $\mathrm{C}(w)$ and $\mathrm{C_I}(w)$ are given by:

$$\mathrm{C}(w) \qquad \{\min 1 \cdot y : yM \geqq w, y \geqq 0\}$$

$$\mathrm{C_I}(w) \qquad \{\min 1 \cdot y : yM \geqq w, y \geqq 0, y \text{ integer}\}.$$

Observe that $\mathrm{C}(w)$ and $\mathrm{C_I}(w)$ have bounded objective value and that these programs are feasible for all $w \geqq 0$ if and only if M has no zero columns; when discussing the IRU property we will assume this to be the case.

Packing and covering problems such as $\mathrm{P}(w)$, $\mathrm{P_I}(w)$, $\mathrm{C}(w)$, $\mathrm{C_I}(w)$ arise naturally in combinatorial optimization (e.g., see [11], [12], [13]). Instances in which IRU or IRD hold have been studied in [20], [14], [21], [18], [19], [3], [4], [1], [2].

In the present paper we establish the IRU and IRD properties for certain classes of matrices arising in the context of polymatroid theory (see [7], [15]) and branching theory (see [9]). The rounding results presented here were strongly motivated by combinatorial packing and covering results for matroids due to Edmonds and Fulker-

* Received by the editors September 19, 1980.

† Solomon Brothers, 1 New York Plaza, New York, New York 10004. The research of this author was partially supported by the National Science Foundation under grant ENG 76-09936.

‡ School of Operations Research/Industrial Engineering, Cornell University, Ithaca, New York 14583, and Institut für Ökonometrie und O. R., Universität Bonn, West Germany. The research of this author was partially supported by the National Science Foundation under grant ENG 76-09936 and Sonderforschungsbereich 21 (DFG).

son (see [10]), especially by the well-known combinatorial min-max theorem of Edmonds [6] on covering the elements of a matroid by its independent sets.

2. Integral decomposition. Denote by R_+^n the nonnegative orthant of Euclidean n-space. A polyhedron $P \subseteq R_+^n$ is called *upper comprehensive* if $y \geqq x \in P$ implies $y \in P$; similarly, P is *lower comprehensive* if $0 \leqq y \leqq x \in P$ implies $y \in P$. Fulkerson's blocking polyhedra (see [11], [12]) are of the form $P = \{x \in R_+^n : Mx \geqq 1\}$ where M is nonnegative; clearly such polyhedra are upper comprehensive. Anti-blocking polyhedra (see [12], [13]) are of the form $P = \{x \in R_+^n : Mx \leqq 1\}$ where M is nonvacuous, nonnegative and has no zero columns; thus anti-blocking polyhedra are nonempty, bounded and lower comprehensive. Given any polyhedron $P \subseteq R_+^n$ and any real number $r \geqq 0$, let rP denote the polyhedron $\{rx : x \in P\}$. The *(integral) decomposition property* holds for P if, for each integer $k \geqq 1$ and each integral vector $x \in kP$, there exist integral vectors $x^i \in P$, $1 \leqq i \leqq k$, for which $x = \sum_{i=1}^k x^i$. The decomposition property has been studied for polyhedra defined by matrices related to network flow problems (see [20], [14], [21]) and polyhedra defined by general totally unimodular constraint matrices (see [19], [4]). We show below that integral decomposition for upper and lower comprehensive polyhedra with integral extreme points is closely related to the IRD and IRU properties for matrices whose rows are given by certain families of integer points within the polyhedra.

Suppose polyhedron $P \subseteq R_+^n$ is upper comprehensive with integral extreme points and let M be the matrix whose rows are the minimal integral points of P. (An integral vector $x \in P$ is a minimal integral point of P if there is no integral vector $y \in P$ with $x \neq y \leqq x$.) It is not difficult to see that M has finitely many rows and that M is nonvacuous and has no rows consisting entirely of zeros if and only if $\emptyset \neq P \subsetneq R_+^n$. Alternatively, suppose $P \subseteq R_+^n$ is a lower comprehensive polyhedron with integral extreme points and list the maximal integral vectors of P as the rows of matrix M. Here M is nonvacuous with finitely many rows if and only if P is nonempty and bounded; thus when M is nonvacuous it follows that M has no zero columns if and only if P has nonempty interior, i.e., $x > 0$ for some $x \in P$.

LEMMA 1. *Let $r \in R_+$ and let $P \subsetneq R_+^n$ be a nonempty polyhedron with integral extreme points.*

(a) *Let the rows of matrix M be the minimal integral points of P. Then* P(w) *has a feasible solution of value r if and only if $x \leqq w$ for some $x \in rP$.*

(b) *Assume further that P is bounded and let the rows of matrix M be the maximal integral points of P. Then* C(w) *has a feasible solution of value r if and only if $x \geqq w$ for some $x \in rP$.*

Proof. (a) If y satisfies $yM \leqq w$, $y \geqq 0$ with $1 \cdot y = r$, then $yM \in rP$ and $yM \leqq w$, so we take $x = yM$. The converse is also clear for $r = 0$, so suppose $r > 0$ and let I index the extreme points of P and J index the extreme rays of P. We have $x \in rP$ and $x \leqq w$. Thus $x/r \in P$ and we may write $x/r = \sum_{i \in I} \lambda_i x^i + \sum_{j \in J} \mu_j z^j$, where x^i are (integral) extreme points of P, $\lambda_i \geqq 0$, $\sum_{i \in I} \lambda_i = 1$, z^j are the extreme rays of P and $\mu_j \geqq 0$. Since P is contained in R_+^n, we must have $z^j \geqq 0$ for each j; thus we have $x/r = \sum_{i \in I} \lambda_i x^i + z$ as above with $z \geqq 0$. Hence $x/r \geqq \sum_{i \in I} \lambda_i x^i$. Furthermore, since the x^i are integral points of P, for each i there exists a row m^i of M so that $x^i \geqq m^i$ (not all the m^i need be distinct). Thus $x/r \geqq \sum_{i \in I} \lambda_i m^i$, and since $x \leqq w$ we obtain that $w \geqq \sum_{i \in I} (r\lambda_i) m^i$. Associating weight $r\lambda_i$ with the component of y corresponding to row m^i of M (and summing such weights when distinct indices of I correspond to the same row of M) and setting the remaining components of y to 0 now gives the desired solution of $P(w)$ of value r.

(b) The proof here is similar to that in part (a). For the necessity, assume $yM \geqq w$, $y \geqq 0$ and $1 \cdot y = r$. Then $x = yM$ satisfies $x \in rP$, $x \geqq w$. For $r = 0$ the converse is clear, so let $r > 0$ and let I index the extreme points of P. Then $x/r \in P$, and so $x/r = \sum_{i \in I} \lambda_i x^i$ with $\sum_{i \in I} \lambda_i = 1$, $\lambda_i \geqq 0$ for each $i \in I$, and each x^i an (integral) extreme point of P. To each vector x^i there corresponds a row m^i of M for which $m^i \geqq x^i$. Thus we have $w \leqq x \leqq \sum_{i \in I} (r\lambda_i) m^i$ and we may use the weights $r\lambda_i$, $i \in I$, as in part (a) to define a solution of $C(w)$ of value r. □

Using Lemma 1 we now establish for the present context an equivalence between the integer rounding properties and the integral decomposition property.

THEOREM 1. (a) *Suppose P is a nonempty upper comprehensive polyhedron with integral extreme points for which $P \subsetneq R_+^n$; let the rows of matrix M be the minimal integral points of P. Then* IRD *holds for M if and only if P satisfies the integral decomposition property.*

(b) *Suppose P is a lower comprehensive polyhedron with integral extreme points which is bounded and has nonempty interior; let the rows of matrix M be the maximal integral points of P. Then* IRU *holds for M if and only if P satisfies the decomposition property.*

Proof. (a) Suppose IRD holds for M and let $w \in kP$, where $k \geqq 1$ is an integer and w is an integral vector. By part (a) of Lemma 1, P(w) has a feasible solution of value k, and thus IRD for M implies that P(w) has an integral solution of value k. Thus there are rows $m^1, \cdots, m^k$ of M (not necessarily all distinct) such that $\sum_{i=1}^k m^i \leqq w$. Now $m^1, \cdots, m^k$ are integral points of P, and since P is upper comprehensive, there are integral points x^i of P, $1 \leqq i \leqq k$, so that $\sum_{i=1}^k x^i = w$. This is the desired decomposition of w.

For the converse, assume that the decomposition property holds for P, and for the nonnegative integral vector w, suppose P(w) has value r^*. If $0 \leqq r^* < 1$, then it is clear that IRD holds for w ($y = 0$ provides a feasible solution to P(w) of value $\lfloor r^* \rfloor = 0$), so assume $r^* \geqq 1$. By Lemma 1, part (a), there exists a vector $x \in r^*P$, $x \leqq w$. Thus $w \in r^*P$, since P upper comprehensive implies that r^*P is also upper comprehensive. Now $\lfloor r^* \rfloor \leqq r^*$, so $r^*P \subseteq \lfloor r^* \rfloor P$; thus we also have $w \in \lfloor r^* \rfloor P$. The decomposition property for P now implies that $w = \sum_{i=1}^{\lfloor r^* \rfloor} x^i$, where the x^i are integral vectors of P. Thus, by the definition of M, there are rows m^i of M such that $m^i \leqq x^i$, for $1 \leqq i \leqq \lfloor r^* \rfloor$, (the m^i need not be distinct). Hence $w \geqq m^1 + \cdots + m^{\lfloor r^* \rfloor}$ determines the desired solution of P(w) of value $\lfloor r^* \rfloor$.

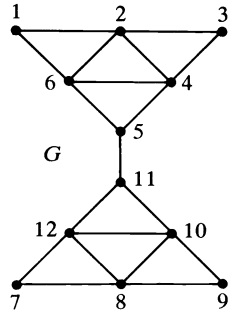
(b) The proof is similar to that given in part (a). □

As an application of Theorem 1, let P be the polyhedron whose extreme points are the (0, 1)-valued incidence vectors of cliques in a graph G and let M have as rows the incidence vectors of all maximal cliques in G. Now P is nonempty and bounded with integral extreme points and it is well known (see [16], [13]) that if G is perfect, then for any nonnegative integral vector w

$$\min \{1 \cdot y : yM \geqq w, y \geqq 0, y \text{ integer}\} = \min \{1 \cdot y : yM \geqq w, y \geqq 0\}.$$

Thus part (b) of Theorem 1 implies that P satisfies the decomposition property. Furthermore, where γ denotes the size of a largest clique in G, decomposition for P implies that the decomposition property also holds for the polyhedron $P' = \{x \in P : 1 \cdot x = \gamma\}$. Let M' be the incidence matrix of maximum cardinality cliques in G. Then it is clear that P' is nonempty and bounded with integer extreme points given by the rows of matrix M'. However, P' is not upper comprehensive, and consequently, IRD need not hold for matrix M'. To see this consider the example below.

Example 1.



$$M' = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ w = \end{array} \begin{array}{c} \begin{array}{cccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \end{array} \\ \left[\begin{array}{cccccccccccc} 1 & 1 & 0 & 0 & 0 & 1 & & & & & & \\ 0 & 1 & 1 & 1 & 0 & 0 & & & & & & \\ 0 & 0 & 0 & 1 & 1 & 1 & & & 0 & & & \\ 0 & 1 & 0 & 1 & 0 & 1 & & & & & & \\ & & & & & & 1 & 1 & 0 & 0 & 0 & 1 \\ & & & 0 & & & 0 & 1 & 1 & 1 & 0 & 0 \\ & & & & & & 0 & 0 & 0 & 1 & 1 & 1 \\ & & & & & & 0 & 1 & 0 & 1 & 0 & 1 \end{array}\right] \\ (1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1) \end{array}$$

Here the solution value of $P_I(w)$ is 2, but $y = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0)$ gives a feasible solution to $P(w)$ of value 3. Thus IRD fails for M'.

The relationship between integer rounding and decomposition given in Theorem 1 is used in the sequel to establish integer rounding in certain instances. A general recursive characterization of the decomposition property may be obtained as follows. Let A be an $m \times n$ rational matrix and let b be a rational m-vector; define the polyhedron $P = \{x \in R_+^n : Ax \leqq b\}$. Now decomposition for the polyhedron P is equivalent to the requirement that for each integral vector $w \in kP$, where k is a positive integer, there must exist an integral vector $x \in P$ for which $(w - x) \in (k-1)P$. That is, the following system must have an integral solution:

$$(1) \qquad Aw - (k-1)b \leqq Ax \leqq b, \qquad 0 \leqq x \leqq w.$$

For certain applications one may show the stronger result that the appropriate polyhedron defined by (1) actually has all integral extreme points (see Corollary 2 of § 3). This would imply the decomposition property, and the approach of Theorem 1 could then be used to establish integer rounding results. However, it is not always the case that the rounding property implies, conversely, that the appropriate polyhedron of the form (1) has all integral extreme points. This is illustrated in the following example.

Example 2. Let matrix M be defined as follows:

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

IRU holds for matrix M. In [5] we show that in order to verify the IRU property for M it suffices to consider the programs $C(w)$, $C_I(w)$ for integral vectors w in the range $0 \leqq w \leqq (2, 2, 2, 1, 1, 1)$. Thus it is tedious, but straightforward, to check that IRU holds for M. From Theorem 1(b) we can then conclude that the polyhedron $P = \{x \in R_+^6 : x \leqq \lambda M \text{ where } \lambda \geqq 0,\ 1 \cdot \lambda = 1\}$ satisfies the decomposition property. Now let $w = (1, 1, 1, 1, 1, 1)$ and $k = 2$. Then it is easy to see that the only *integral* vectors x satisfying $x \in P$, $(w - x) \in P$ are rows 4 and 5 of matrix M. However, the vector $x = (1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0)$ clearly satisfies $x \in P$, $(w - x) \in P$, but this vector is *not* a convex combination of rows 4 and 5 of M. Hence the bounded, nonempty polyhedron $\{x : x \in P, (w - x) \in P\}$ cannot have only integral extreme points.

3. Polymatroid optimization. In [7] (see also [15]) Edmonds has characterized *polymatroids* as polyhedra of the form

$$P(E, f) = \{x \in R_+^n : x(S) \leqq f(S), S \subseteq E\},$$

where $E = \{1, 2, \cdots, n\}$, $x(S)$ denotes the quantity $\sum_{j \in S} x_j$ and f is a real-valued function on subsets of E satisfying three conditions:

(i) $f(S) \geqq 0$, $S \subseteq E$ (nonnegativity),
(ii) $R \subseteq S \Rightarrow f(R) \leqq f(S)$, $R, S \subseteq E$ (monotonicity),
(iii) $f(R \cup S) + f(R \cap S) \leqq f(R) + f(S)$, $R, S \subseteq E$ (submodularity).

When f is also integer-valued, $P(E, f)$ is an *integral polymatroid.* A well-known instance of an integral polymatroid is obtained by taking f as the rank function of a matroid defined on E; then $P(E, f)$ is the familiar "matroid polyhedron", whose extreme points are the incidence vectors of independent sets in that matroid (see [8]). It is also the case that general integral polymatroids have integral extreme points (see [7], [15]).

We now show that integral polymatroids satisfy the decomposition property. Our development of this result, culminating in Theorem 2, is algebraic and it follows closely the development in [7] used in obtaining the polymatroid intersection theorem.

A family F of subsets of E satisfies *property* $(*)$ when

$$(*) \qquad R \cap S \neq \varnothing \Rightarrow (R \cap S) \in F, \qquad R, S \in F.$$

The following three lemmas on property$(*)$ are taken directly from [7].

LEMMA 2. *Let F be a family of subsets of E satisfying property $(*)$ and let $M = (m_{ij})$ be the incidence matrix of F with E, i.e., where $F = \{T_1, T_2, \cdots, T_m\}$ we have $m_{ij} = 1$ for $j \in T_i$ and $m_{ij} = 0$ otherwise. From M one can obtain the incidence matrix of a family of disjoint subsets of E by subtracting certain rows of M from others.*

LEMMA 3. *Let F_1, F_2 be two families of subsets of E each satisfying property $(*)$ and let $M = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}$, where M_i is the incidence matrix of F_i with E, for $i = 1, 2$. From M one can obtain a totally unimodular matrix by subtracting certain rows of M from others.*

LEMMA 4. *Let $x^* \in \{x \in R^n : x(T) \leqq f(T), T \subseteq E\}$, where f is a submodular real-valued function. Then $F = \{T \subseteq E : x^*(T) = f(T)\}$ satisfies property $(*)$.*

PROOF OF LEMMAS 2–4. One proves Lemma 2 by the recursive subtraction of a minimal row of M from the remaining rows of M which dominate it. Lemma 3 then follows by applying Lemma 2 to both M_1 and M_2, which transforms M into the incidence matrix of a bipartite graph. Lemma 4 follows from the following relations, where $R, S \in F$ and $R \cap S \neq \varnothing$:

$$\begin{aligned} f(R \cap S) &\leqq f(R) + f(S) - f(R \cup S) \\ &\leqq x^*(R) + x^*(S) - x^*(R \cup S) = x^*(R \cap S) \leqq f(R \cap S). \end{aligned}$$

□

When k is a positive integer, $z \in R^n$ and f is a submodular real valued function on subsets of E, the function defined by $(k-1)f(T) - z(T)$, $T \subseteq E$ is also submodular. Thus Lemma 4 implies

COROLLARY 1. *Let $x^* \in \{x \in R^n : z(T) - (k-1)f(T) \leqq x(T), T \subseteq E\}$, where f is real-valued and submodular, $z \in R^n$ and k is a positive integer. Then $F = \{T \subseteq E : z(T) - (k-1)f(T) = x^*(T)\}$ satisfies property $(*)$.*

A further family of subsets of E satisfying property$(*)$ is given by

LEMMA 5. *Let $x^* \in \{x \in R^n : x(T) \leqq \min\{f(T), z(T)\}, T \subseteq E\}$, where f is real-valued and submodular and $z \in R^n$. Then $F = \{T \subseteq E : x^*(T) = \min\{f(T), z(T)\}\}$ satisfies property $(*)$.*

Proof. Suppose $R, S \in F$ with $R \cap S \neq \emptyset$. We consider 3 cases.

(i) $x^*(R) = f(R)$ and $x^*(S) = f(S)$. Here Lemma 4 applies.

(ii) $x^*(R) = z(R)$. Now $x^*(R') \leqq z(R')$ for all $R' \subseteq R$ and hence $x^*(R') = z(R')$ for all $R' \subseteq R$. Thus $x^*(R \cap S) = z(R \cap S) \geqq \min\{f(R \cap S), z(R \cap S)\}$. But by assumption $x^*(R \cap S) \leqq \min\{f(R \cap S), z(R \cap S)\}$. Thus we have $x^*(R \cap S) = \min\{f(R \cap S), z(R \cap S)\}$, as required.

(iii) $x^*(S) = z(S)$. The argument is similar to that in case (ii). □

THEOREM 2. *Let $P(E, f)$ be an integral polymatroid and let $z \in kP(E, f)$ be an integral vector, where k is a positive integer. Then there exist integral vectors $x^1, x^2, \cdots, x^k \in P(E, f)$ for which $z = x^1 + x^2 + \cdots + x^k$.*

Proof. The proof is by induction on k; since the result clearly holds when $k = 1$, suppose it true for $1, 2, \cdots, k-1$. To establish the result for k it suffices to determine an integral vector $x \in P(E, f)$ so that $(z - x) \in (k-1)P(E, f)$; such an x would serve as x^k and then induction could be applied to the integral remainder $(z - x) \in (k-1)P(E, f)$. Hence we seek an integral vector in the polyhedron

$$P = \{x \in R_+^n : z(T) - (k-1)f(T) \leqq x(T) \leqq \min\{f(T), z(T)\},\ T \subseteq E\}.$$

Note that $(z/k) \in P$, so that $P \neq \emptyset$. Thus we complete the proof by showing that P has integral extreme points.

Suppose x^* is an extreme point of P and define $F_0 = \{\{j\} : x_j^* = 0\}$, $F_1 = \{T \subseteq E : z(T) - (k-1)f(T) = x^*(T)\}$, $F_2 = \{T \subseteq E : x^*(T) = \min\{f(T), z(T)\}\}$. Let M_i be the incidence matrix of F_i, $i = 1, 2$, and let J be the incidence matrix of F_0. Since x^* is an extreme point of P, it is the unique solution to the equality system

$$\begin{bmatrix} M_1 \\ M_2 \\ J \end{bmatrix} x = \begin{bmatrix} b^1 \\ b^2 \\ 0 \end{bmatrix},$$

where the value of the component of vector b^1 corresponding to subset $T \in F_1$ is $z(T) - (k-1)f(T)$ and for $T \in F_2$ the corresponding component of b^2 has value $\min\{f(T), z(T)\}$.

By Corollary 1 and Lemma 5 we see that F_1 and F_2 satisfy the hypothesis of Lemma 3. Thus x^* is also the unique solution to

$$\begin{bmatrix} N_1 \\ N_2 \\ J \end{bmatrix} x = \begin{bmatrix} c^1 \\ c^2 \\ 0 \end{bmatrix},$$

where N_i and c^i are obtained from M_i and b^i, $i = 1, 2$, by subtracting certain rows from others, and $\left[\begin{smallmatrix} N_1 \\ N_2 \end{smallmatrix}\right]$ is a totally unimodular matrix. Thus

$$\begin{bmatrix} N_1 \\ N_2 \\ J \end{bmatrix}$$

is also totally unimodular. Since k, z and f are integral, it follows that b^1, b^2 are integral vectors and hence c^1, c^2 are integral vectors. Thus x^* is also integral. □

A byproduct of the proof of Theorem 2 is

COROLLARY 2. *Suppose f is an integer-valued submodular function on subsets of $E = \{1, 2, \cdots, n\}$, z is an integral n-vector and k is a positive integer. Then each extreme point of the following polyhedron is integral:*

$$\{x \in R_+^n : z(T) - (k-1)f(T) \leqq x(T) \leqq \min\{f(T), z(T)\},\ T \subseteq E\}.$$

A generalization of Theorem 2 to the context in which k integral polymatroids are defined on E was established independently (and earlier) by Giles (see [15, Thm. 4.6.6]). We can now use the integral decomposition of Theorem 2 to establish integer rounding results for integral polymatroids.

THEOREM 3. *Suppose $P(E,f)$ is an integral polymatroid with $\emptyset \neq P(E,f) \neq \{0\}$ and let matrix M have as rows the maximal integral vectors in $P(E,f)$.*

(a) *The* IRD *property holds for M.*

(b) *The* IRU *property holds for M, provided $f(\{i\}) > 0$ for all $i \in E$.*

Proof. Clearly $P(E,f)$ is nonempty and lower comprehensive. Since f is integral, $P(E,f)$ is bounded and has integral extreme points; furthermore, when $f(\{i\}) > 0$ for $i \in E$, $P(E,f)$ has nonempty interior. Thus Theorem 2 and part (b) of Theorem 1 combine to establish (b). To prove part (a), let $w \geqq 0$ be integral and suppose y^* solves $P(w)$. When $\lfloor 1 \cdot y^* \rfloor = 0$, the result is clear, so suppose $\lfloor 1 \cdot y^* \rfloor \geqq 1$ and denote by P the convex hull of the rows of M. The maximal integral vectors of $P(E,f)$, i.e., the rows of M, are those integral vectors $x \in P(E,f)$ for which $x(E) = f(E)$ (see [7]). Thus $P = P(E,f) \cap \{x : x(E) = f(E)\}$. Applying Corollary 2 to the integral polymatroid $\lfloor 1 \cdot y^* \rfloor P(E,f)$ with k large and $z = w$ shows that $\lfloor 1 \cdot y^* \rfloor P(E,f) \cap \{x : x \leqq w\}$ has integral extreme points (which may also be seen directly from the polymatroid intersection theorem). By intersecting with the supporting hyperplane $\{x : x(E) = \lfloor 1 \cdot y^* \rfloor f(E)\}$, we obtain that the polyhedron $\lfloor 1 \cdot y^* \rfloor P \cap \{x : x \leqq w\}$ has integral extreme points. The latter polyhedron is nonempty, as it contains the point $(\lfloor 1 \cdot y^* \rfloor / 1 \cdot y^*) y^* M$. Thus there exists an integral vector $z \in \lfloor 1 \cdot y^* \rfloor P \subseteq \lfloor 1 \cdot y^* \rfloor P(E,f)$, $0 \leqq z \leqq w$. Theorem 2 shows that there exist integral vectors $x^i \in P(E,f)$, $1 \leqq i \leqq \lfloor 1 \cdot y^* \rfloor$, with $z = x^1 + x^2 + \cdots + x^{\lfloor 1 \cdot y^* \rfloor}$. Since $z \in \lfloor 1 \cdot y^* \rfloor P$, it follows that each $x^i \in P$. Thus the vectors x^i, $1 \leqq i \leqq \lfloor 1 \cdot y^* \rfloor$, are rows of M, and hence they determine a solution to $P_I(w)$ of value $\lfloor 1 \cdot y^* \rfloor$. □

Theorem 3 may be used to derive combinatorial min-max and max-min theorems involving integer rounding in the following way. Fulkerson (see [12], [13]) has shown that if M is a nonnegative matrix without zero columns and A is an anti-blocking matrix for M, then the following min-max relation holds for all $w \geqq 0$:

$$\min \{1 \cdot y : yM \geqq w,\ y \geqq 0\} = \max \{w \cdot \alpha : \text{is a row of } A\}.$$

Now suppose, as above, that the rows of M are the maximal integral vectors in an integral polymatroid $P(E,f)$ which is *loopless*, i.e., $f(T) > 0$ for all $\emptyset \neq T \subseteq E$. Then M has no zero columns and the work of Edmonds (see [7], [8]) shows that an anti-blocking matrix for M is given by matrix A with rows $\{1_T / f(T) : \emptyset \neq T \subseteq E\}$, where 1_T is the incidence vector of subset T. Thus combining the min-max relation above with part (b) of Theorem 3 yields the following integral opimization result: for any integral $w \geqq 0$,

$$\min \{1 \cdot y : yM \geqq w,\ y \geqq 0,\ y \text{ integral}\} = \max \{\lceil w(T)/t(T) \rceil : \emptyset \neq T \subseteq E\}.$$

When f is the rank function of a matroid on E and $w = 1_E$, this becomes the well-known theorem of Edmonds [6] that the smallest number of independent sets in a matroid required to cover its elements is equal to $\max \{\lceil |T| / f(T) \rceil : \emptyset \neq T \subseteq E\}$. Combinatorial max-min results may be similarly derived by combining the max-min equality for blocking pairs of matrices [11], [12], the matroid and polymatroid results of [7], [10] and part (a) of Theorem 3.

4. Branchings. A *branching* in a directed graph is a subgraph which is a forest (i.e., acyclic) no two of whose edges are directed toward the same vertex. Thus the

branchings of G are the common independent sets of two matroids defined on the edges of G: the *forest* matroid whose independent sets are given by acyclic edge sets in the undirected graph underlying G and the *partition* matroid whose independent sets are given by sets of edges of G directed at different vertices of G. Thus, where f_1 and f_2 denote the respective rank functions of these two matroids, it follows from [7] that for graph G with edge set E, the polyhedron

$$P(G) = \{x \in R_+^{|E|} : x(S) \leqq \min\{f_1(S), f_2(S)\},\ S \subseteq E\}$$

has extreme points which are precisely the incidence vectors of branchings in G.

A branching in G is said to be *rooted* at a vertex v of G if each vertex of G except for v has a branching edge directed toward it. In order to establish rounding results for branchings, we first prove a decomposition theorem for $P(G)$ using the following well-known result on branchings. We assume throughout this section that G has a nonempty edge set.

THEOREM 4 (Edmonds [9]). *Let G be a directed graph with vertex set V and edge set E; for $X \subseteq V$ let $\bar{X} = V \backslash X$ and let $(X, \bar{X}) = \{e = (u, w) \in E:\ u \in X,\ w \in \bar{X}\}$. Then the maximum number of edge-disjoint branchings of G rooted at $v \in V$ equals the minimum of $|(X, \bar{X})|$ taken over all X with $v \in X \subsetneq V$.*

We will also require the following result on polymatroid intersection.

THEOREM 5 (McDiarmid [17]). *Suppose p and q are integer scalars, u and w are nonnegative integral n-vectors and $P(E, f_1)$ and $P(E, f_2)$ are integral polymatroids on $E = \{1, 2, \cdots, n\}$. Then the following polyhedron is the convex hull of its integral elements:*

$$P(E, f_1) \cap P(E, f_2) \cap \{x \in R^n : u \leqq x \leqq w;\ p \leqq x(E) \leqq q\}.$$

THEOREM 6. *Let G be a directed graph and let $z \in kP(G)$ be an integral vector, where k is a positive integer and $P(G)$ is as defined above. Then there exist integral vectors $x^1, x^2, \cdots, x^k \in P(G)$ for which $z = x^1 + x^2 + \cdots + x^k$.*

Proof[1]. Suppose G has vertex set V and edge set $E = \{e_1, e_2, \cdots, e_{|E|}\}$. From G we construct the graph G^* with vertices V^* and edges E^* as follows. Define $V^* = V \cup \{v^*\}$ where $v^* \notin V$. For each $e_i = (u, w) \in E$, E^* contains z_i copies of the edge (u, w), and for each $v \in V$, E^* contains $k - \sum(z_i : e_i = (u, v)$ for some $u \in V) \geqq 0$ copies of the edge (v^*, v), where nonnegativity follows because $(z/k) \in P(G)$ and the extreme points of $P(G)$ correspond to branchings in G.

Let $v^* \in X \subsetneq V^*$. By Theorem 2, z is the sum of k incidence vectors of forests in G and so G^* has at most $k(|\bar{X}| - 1)$ edges with both endpoints in $\bar{X} = V^* \backslash X$. But since each vertex in $\bar{X}$ has exactly k edges of G^* directed toward it, we must have $|(X, \bar{X})| \geqq k$. Thus by Theorem 4 G^* contains k edge-disjoint branchings of G^* rooted at v^*. Restricting these branchings in G^* to $V^* \backslash \{v^*\}$ determines k branchings of G whose incidence vectors sum to z.

From Theorem 6 and part (b) of Theorem 1 we obtain

THEOREM 7. *Let matrix M have as rows the incidence vectors of maximal branchings in the directed graph G. Then the* IRU *property holds for M.*

Now let matrix M have as rows the *maximum* cardinality branchings of G and denote by $P'(G)$ the polyhedron which is the convex hull of the rows of M. Since $P(G)$ satisfies the decomposition property, it is not difficult to see that $P'(G)$ also

[1] The authors are indebted to Rick Giles for providing this proof.

satisfies the decomposition property. Using this fact we can establish integer rounding results for $P'(G)$, though $P'(G)$ is neither upper nor lower comprehensive.

THEOREM 8. *Let matrix M have as rows the incidence vectors of maximum cardinality branchings in the directed graph G.*

(a) *The* IRD *property holds for M.*

(b) *The* IRU *property holds for M, provided M has no zero columns, i.e., each edge of G appears in some branching of maximum cardinality.*

Proof. We prove part (a); the proof of (b) is entirely similar. Let $w \geqq 0$ be a given integral vector and suppose y^* solves $P(w)$. For $0 \leqq 1 \cdot y^* < 1$, the result is clear, so suppose $\lfloor 1 \cdot y^* \rfloor \geqq 1$. By Lemma 1(a) there is a vector $y \in (1 \cdot y^*)P'(G)$ such that $y \leqq w$, and so there exists a vector $z \in \lfloor 1 \cdot y^* \rfloor P'(G)$ with $z \leqq w$. By Theorem 5 we may assume that z is an integral vector, and since the decomposition property holds for $P'(G)$, we may write $z = x^1 + x^2 + \cdots + x^{\lfloor 1 \cdot y^* \rfloor}$, where each x^i, $1 \leqq i \leqq \lfloor 1 \cdot y^* \rfloor$, is an integral vector in $P'(G)$. This integral decomposition of z determines the desired solution to $P_I(w)$ of value $\lfloor 1 \cdot y^* \rfloor$. □

REFERENCES

[1] J. J. BARTHOLDI III AND H. D. RATLIFF, *Unnetworks, with applications to idle time scheduling*, Management Sci., 24 (1978), pp. 850–858.

[2] J. J. BARTHOLDI III, J. B. ORLIN AND H. D. RATLIFF, *Cyclic scheduling via integer programs with circular ones*, Oper. Res., 28 (1980) pp. 1074–1085.

[3] S. BAUM, *Integral near-optimal solutions to certain classes of linear programming problems*, Ph.D. Thesis, Tech. Rep. 360, School of Operations Research and Industrial Engineering, Cornell Univ., Ithaca, NY, 1977.

[4] S. BAUM AND L. E. TROTTER, JR., *Integer rounding and polyhedral decomposition for totally unimodular systems*, in Arbeitstagung über Operations Research und Optimierung, R. Henn, B. Korte and W. Oettli, eds., Springer-Verlag, Berlin-Heidelberg-New York, 1978, pp. 15–23.

[5] ———, *Finite checkability for integer rounding properties in combinatorial programming problems*, Universität Bonn, Institut für Ökonometrie und Operations Research, Rep. 78109-OR, 1978, to appear in Math. Programming.

[6] J. EDMONDS, *Minimum partition of a matroid into independent subsets*, J. Res. Nat. Bur. Standards Sect. B., 69 (1965), pp. 67–72.

[7] ———, *Submodular functions, matroids, and certain polyhedra*, in Combinatorial Structures and Their Applications, R. Guy, H. Hanani, N. Sauer and J. Schonheim, eds., Gordon and Breach, New York, 1970, pp. 69–87.

[8] ———, *Matroids and the greedy algorithm*, Math. Programming, 1 (1971), pp. 127–136.

[9] ———, *Edge-disjoint branchings*, in Combinatorial Algorithms, R. Rustin, ed., Algorithmics Press, New York, 1972, pp. 91–96.

[10] J. EDMONDS AND D. R. FULKERSON, *Transversals and matroid partition*, J. Res. Nat. Bur. Standards Sect. B, 69 (1965), pp. 147–153.

[11] D. R. FULKERSON, *Blocking polyhedra*, in Graph Theory and Its Applications, B. Harris, ed., Academic Press, New York, 1970, pp. 93–111.

[12] ———, *Blocking and anti-blocking pairs of polyhedra*, Math. Programming, 1 (1971), pp. 168–194.

[13] ———, *Anti-blocking polyhedra*, J. Combin. Theory, 12 (1972), pp. 50–71.

[14] D. R. FULKERSON AND D. B. WEINBERGER, *Blocking pairs of polyhedra arising from network flows*, J. Combin. Theory, 18 (1975), pp. 265–283.

[15] F. R. GILES, *Submodular functions, graphs and integer polyhedra*, Ph.D. Thesis, Dept. of Combinatorics and Optimization, Univ. of Waterloo, Waterloo, Ontario, 1975.

[16] L. LOVÁSZ, *Normal hypergraphs and the perfect graph conjecture*, Discrete Math., 2 (1972), pp. 253–267.

[17] C. J. H. MCDIARMID, *Blocking, anti-blocking, and pairs of matroids and polymatroids*, J. Combin. Theory Ser. B, 25 (1978), pp. 313–325.

[18] ———, *On pairs of strongly-base orderable matroids*, Tech. Rep. 283, School of Operations Research and Industrial Engineering, Cornell Univ. Ithaca, NY, 1973.

[19] L. E. TROTTER, JR. AND D. B. WEINBERGER, *Symmetric blocking and anti-blocking relations for generalized circulations*, Math. Programming Stud. 8 (1978), pp. 141–158.

[20] D. B. WEINBERGER, *Investigations in the theory of blocking pairs of polyhedra*, Ph.D. Thesis, Tech. Rep. 190, School of Operations Research and Industrial Engineering, Cornell Univ., Ithaca, NY, 1973.

[21] ———, *Network flows, minimum coverings, and the four-color conjecture*, Oper. Res., 24 (1976), pp. 272–290.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 4, December 1981

0196-5212/81/0204-0008 $01.00/0

HYPERGEOMETRIC AND GENERALIZED HYPERGEOMETRIC GROUP TESTING*

F. K. HWANG,† TIEN TAI SONG‡ AND DING ZHU DU‡

Abstract. We consider two group testing problems involving a set of n items. In the first problem *extactly d* defectives, and in the second problem *at most d* defectives, are distributed arbitrarily in the set. We show that any procedure for identifying all defectives in the first problem can be easily adapted to the second problem, with an increase of at most one in the maximum number of tests required. Some related problems are also described.

1. Introduction. In a group testing problem we are concerned with a finite population P of n items each of which can be classified either as good or defective. A *group test* is a simultaneous test on an arbitrary subset $X \subseteq P$ with two possible outcomes: X is *pure* if all items in X are good, X is *contaminated* otherwise. The group testing problem is to find all the defectives in the population (hence all items are classified) by means of a sequence of group tests, with the aim of minimizing the number of such tests.

When the number of defectives d in the population is known exactly and the defectives are distributed arbitrarily in the population, the problem is known as the *hypergeometric group testing problem* and is denoted by (n, d). Due to the mathematical simplicity of its assumptions, the hypergeometric group testing problem has attracted a great deal of attention from research workers. On the other hand, in most practical situations the exact number of defectives is rarely deducible, though an upper bound on it is usually available. When our prior knowledge consists of an upper bound $\bar{d}$ instead of the exact number of defectives, the problem is known as the *generalized hypergeometric group testing problem* and is denoted by $(n, \bar{d})$.

It is desirable to establish a general method such that any procedure for the hypergeometric group testing problem can be modified to apply to the generalized version. Such an attempt was made in [1], which showed that if we only consider procedures of a special but important class called "nested," then procedures for one problem can be immediately translated to procedures for the other problem with essentially no effect on the maximum number of tests required. In this paper we remove the restriction to nested procedures and prove a similar result. Some related open problems are assembled in the last section.

2. Some preliminary remarks. A group testing procedure can be represented by a binary tree (see [1] for terminology) where each internal node is associated with a group test and its two outlinks are associated with the two possible outcomes of the test. The *test history* H_v at node v is the set of tests and outcomes associated with the nodes and links on the path from the root to v, excluding v itself. Let D, called the *defective set*, denote the set of defectives in the population. In the (n, d) problem, D can be any d-subset of the n items, and in the $(n, \bar{d})$ problem D can be any k-subset with $0 \leqq k \leqq d$. Suppose S is the set of possible solutions for D and r is a procedure for S. We associate S with the root of r. For any other node v of r, we define S_v to be the set of elements in S which are *consistent* with H_v. Namely, let $s \in S_v$ and t be a group tested in H_v. Then $s \cap t = \varnothing$ if and only if t is pure. Thus if v is an internal

* Received by the editors February 5, 1981, and in final form March 23, 1981.
† Bell Laboratories, Murray Hill, New Jersey 07974.
‡ Institute of Applied Mathematics, Academia Sinica, Beijing, China.

node, the test at v partitions S_v into two smaller sets consistent with the two possible outcomes respectively. If v is a terminal node, then S_v consists of a single element, which we denote by s_v.

The number of tests in H_v is clearly identical to the length of the path from the root to v. Let $M_r(S)$ denote the maximum path length over all terminal nodes in the procedure r where S is the initial set of possibilities for D given in the problem. Define $M(S) = \min_r M_r(S)$, called the *minimax number* for S.

LEMMA 1. *Suppose* $S \subset S'$. *Then* $M(S) \leqq M(S')$.

Proof. Any procedure for S' is clearly also a procedure for S, with certain tests possibly being redundant. Lemma 1 follows immediately. □

LEMMA 2. *Let* S *consist of all* $(n-1)$*-subsets of an* n*-set. Then* $M(S) = n-1$.

Proof. There is no need to test any group of cardinality greater than one since the outcome must be contaminated. Therefore the optimal algorithm is that which tests the n items one by one, which requires $n-1$ tests when the outcomes of the first $n-2$ tests are all contaminated. □

3. The main results. Let r denote a procedure for the (n, d) problem. Let v be a terminal node of r and $S_v = \{s_v\}$. Partition the d items in s_v into two categories, the *fixed items* and the *free items*. $\omega \in s_v$ is a fixed item if there exists a group tested in H_v such that the group does not contain any other element of s_v; otherwise, ω is a free item. A free item is identified as defective through the identification of $n-d$ good items.

LEMMA 3. *Suppose that* v *is a terminal node with* $f \geqq 1$ *free items. Let* u *be the* brother *node of* v, *i.e.*, u *and* v *are associated with the two outcomes of a test. Then* u *is the root of a subtree whose maximum path length is at least* $f-1$.

Proof. Let t denote the last group tested before v. Since whenever $n-d$ good items are identified the (n, d) problem is necessarily solved, free items can be identified only at the last test. Therefore $f \geqq 1$ implies that v corresponds to the pure outcome of t. Hence t cannot contain any item of s_v.

Let s denote a free item of s_v and $\omega \in t$. Then $s_v - \{s\} \cup \{\omega\} \in S_u$. To see this, note that any test group containing s must contain another element of s_v or s would be a fixed item. Therefore changing s from defective to good does not change the outcome of any test in H_v. Furthermore $s_v - \{s\} \cup \{\omega\}$ is consistent with the contaminated outcome of t, hence it is in S_u.

Let S' denote the set $\{s_v - \{s\} \cup \{\omega\}: s \in s_v \text{ and } s \text{ is free}\}$. Then $M(S_u) \geqq M(S') = |S'| - 1 = f-1$ by Lemmas 1 and 2. □

THEOREM. *For each procedure* r *for the* (n, d) *problem, there exists a procedure* r' *for the* $(n, \bar{d})$ *problem such that*

$$M_r(n, d) + 1 \geqq M_{r'}(n, \bar{d}).$$

Proof. Let r' be obtained from r by adding a subtree T_v to each terminal node v having a positive number of free items. T_v is the tree obtained by testing the free items one by one. Since free items are the only items at v whose states are uncertain when we change from (n, d) to $(n, \bar{d})$, r' is a procedure for the $(n, \bar{d})$ problem. From Lemma 3, the brother node of v is the root of a subtree with maximum path length at least $f-1$ where f is the number of free items of s_v. The theorem follows immediately. □

COROLLARY. $M(n, d) + 1 \geqq M(n, \bar{d}) \geqq M(n+1, d)$.

Proof. The first inequality follows from the theorem. The second inequality follows from the observation that the $(n+1, d)$ problem can be solved by any procedure for

the $(n, \bar{d})$ problem, provided one of the $n+1$ items is put aside. But the nature of the item put aside can be deduced with certainty once the natures of the other n items are known. □

4. Some concluding remarks. We here list a number of related problems.

(i) *The determination of* $M(n, 2)$. That $M(n, 1) = \lceil \log_2 n \rceil$, where $\lceil x \rceil$ denotes the smallest integer not less than x, is trivial. But one step further leads to the incredibly hard $M(n, 2)$ problem (see [3] for a good algorithm). A conjecture related to this problem was finally solved in [1], [2]: "Given two disjoint populations of m and n items each containing exactly one defective, $\lceil \log_2 mn \rceil$ tests suffice as the maximum number of tests to find the two defectives."

(ii) *The relation between n and d such that* $M(n, d) = n - 1$. This is a longstanding problem with but little progress. The conjecture that $M(n, d) = n - 1$ for $n \leqq 2d + 1$ was floating around in 1970 and proved in [6]. It took a ten-year span for the next result [5], $M(n, d) = n - 1$ for $n \leqq \lfloor (5d + 1)/2 \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer not exceeding x. Recently [4], this result has been further improved to $M(n, d) = n - 1$ for $n \leqq \lfloor 21d/8 \rfloor$. On the other hand, it was proved that $M(n, d) < n - 1$ for $n \geqq 3d$ and conjectured that $M(n, d) = n - 1$ for $n < 3d$ in [5].

(iii) *A monotonicity property in binomial group testing.* The setting of the problem is slightly different from the other problems in that the population consists of n stochastically independent items each with probability p of being defective. When probabilities are involved, the maximum number of tests is no longer a suitable criterion. Instead, we are concerned with the expected number of tests. Let $E(n, p)$ denote the minimum expected number of tests for such a population. The conjecture is that $E(n, p)$ is monotone increasing in p for $0 \leqq p < 1$.

REFERENCES

[1] G. J. CHANG AND F. K. HWANG, *A group testing problem*, this Journal, 1 (1980), pp. 21–24.
[2] ———, *A group testing problem on two disjoint sets*, this Journal, 2 (1981), pp. 35–38.
[3] G. J. CHANG, F. K. HWANG AND S. LIN, *Group testing with two defectives*, to appear.
[4] D. Z. DU AND F. K. HWANG, *Minimizing* $\binom{(m+n)k+l}{mk+\lfloor (l+1)m/(m+n) \rfloor} \lambda^k$ *over k*, to appear.
[5] M. C. HU, F. K. HWANG AND J. K. WANG, *A boundary problem for group testing*, this Journal, 2 (1981), pp. 81–87.
[6] HWANG, F. K., *A minimax procedure on group testing problems*, Tamkang J. Math., 2 (1971), pp. 39–44.
[7] ———, *A note on hypergeometric group testing*, SIAM J. Appl. Math., 34 (1978), pp. 371–375.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 4, December 1981

0196-5212/81/0204-0009 $01.00/0

ACYCLIC DIGRAPHS, YOUNG TABLEAUX AND NILPOTENT MATRICES*

EMDEN R. GANSNER†

Abstract. A nilpotent matrix is associated with an acyclic digraph in such a way that the Jordan invariants of the matrix correspond to the maximum size of certain families of paths in the digraph. This allows one to associate an integer partition and a standard Young tableau with the digraph, extending the Robinson-Schensted map on permutations. The associated partition is characterized using matrices whose rows are paths in the digraph. This leads to a proof of a conjecture of Greene concerning the entries in the associated Young tableau. When the digraph is transitive, a second characterization is given for the partition. Most of the arguments used are algebraic in nature.

1. Introduction. Recently, a number of combinatorial properties of posets have been proved using the tools of linear algebra [17], [20], [21], [24]. One looks at the vector space $\mathbf{V}$ over, say, the complex field, freely generated by the vertices of the poset, and finds that certain properties of the poset translate into properties of certain linear maps of $\mathbf{V}$ into itself, which can be verified algebraically. The maps thus involved are essentially elements in the incidence algebra [18] of the poset.

The present paper continues in this spirit, while at the same time generalizing the focus from posets to acyclic digraphs. Given such a digraph Γ, we associate with it a nilpotent matrix, and show that the invariants of the Jordan canonical form of the matrix correspond to the maximum size of certain families of paths in the graph.

Using this result, we see how to associate a partition Δ' of an integer and, more, a standard Young tableau Y with Γ, thereby extending and rederiving results of Fomin [5] and Greene [9]. In addition, we note a generalization of a theorem of Greene and Kleitman [8] concerning k-saturated partitions of posets to acyclic digraphs. This is used to follow Greene's lead [10] and characterize Δ' in terms of matrices whose rows are paths in Γ. We are then able to prove a conjecture of Greene [11] concerning the entries in Y. As a corollary to this characterization, we find that Y gives us a generalization of the well-known Schensted correspondence [22].

Using strictly combinatorial arguments, but directed by the above-mentioned characterization of Δ', we obtain a second characterization of Δ' when Γ is a poset, based on matrices whose rows are antichains in Γ. It is also noted that an obvious, second characterization of Y does not work.

Throughout this paper, we assume a basic familiarity with linear algebra and fields (rank, nullity, Jordan canonical form, algebraic independence), graphs, matroids, and posets, as can be derived from such texts as [1], [12], [13], [25]. More detailed results will be cited as needed.

2. Generic matrices, k-paths, and partitions. Let Γ be a finite acyclic digraph (or A-digraph) on g vertices, with the vertices labeled from 1 to g. A vertex is a *sink* if no edge proceeds from it; a vertex is a *source* if no edge leads into it. The ordered pair (i, j) represents an edge from vertex i to vertex j.

A *path* in Γ is a sequence of vertices $i_1, i_2, \cdots, i_n$ such that (i_j, i_{j+1}) is an edge in Γ for $1 \leqq j \leqq n-1$. We do allow a path to consist of a single vertex. A *k-path* in Γ is a subset of the vertices that can be partitioned into k or fewer disjoint paths. We let $\hat{d}_k = \hat{d}_k(\Gamma)$ be the largest cardinality of a k-path in Γ, with $\hat{d}_0 = 0$ by convention. In

* Received by the editors December 8, 1980.
† Bell Laboratories, Murray Hill, New Jersey 07974.

addition, we define $\hat{\Delta}_k = \hat{\Delta}_k(\Gamma) = \hat{d}_k - \hat{d}_{k-1}$ for $k \geqq 1$. Let $\hat{\Delta}$ denote the infinite sequence $\hat{\Delta}_1, \hat{\Delta}_2, \hat{\Delta}_3, \cdots$. Clearly, we have $\hat{d}_k \leqq \hat{d}_{k+1}$, so the $\hat{\Delta}_k$'s are nonnegative.

Our main tool throughout this paper will be the ability to give a linear algebraic interpretation to the $\hat{\Delta}_k$'s. To this end, we define a *generic matrix* as a square matrix with entries from the complex numbers $\mathbb{C}$ whose nonzero entries are all algebraically independent over the rational numbers $\mathbb{Q}$. Given an A-digraph Γ, with g vertices, define a $g \times g$ matrix $M_\Gamma = (m_{ij})$ such that $m_{ij} = 0$ if (i, j) is not an edge in Γ and such that the rest of the entries of M_Γ are complex numbers algebraically independent over $\mathbb{Q}$. Since M_Γ is generic, we call it a *generic matrix of* Γ. And, since Γ is acyclic, M_Γ is nilpotent. Conversely, it is easy to see that, given any generic, nilpotent $g \times g$ matrix, it corresponds in the above fashion with a unique labeled A-digraph on g vertices. Note also that any nilpotent generic matrix has only 0's on the main diagonal.

For any nilpotent matrix M, its Jordan canonical form, or, identically, its rational canonical form, will consist of, say, s Jordan blocks of sizes $n_1 \geqq n_2 \geqq \cdots \geqq n_s > 0$, with 0's on their main diagonals. For convenience, let $n_k = 0$ for $k > s$. The n_k's are the *invariants* of M. We now come to the result (obtained independently and through different techniques by Saks [21, Thm. 6.2]) that will allow us to use the machinery of linear algebra in our investigations.

THEOREM 2.1. *Let Γ be an A-digraph. Let M_Γ be a generic matrix of Γ with invariants n_k, $k \geqq 1$. Then $n_k = \hat{\Delta}_k(\Gamma)$ for all $k \geqq 1$.*

To prove this, we will first need a good handle on the canonical form of a nilpotent matrix. If H is a matrix whose entries are polynomials in the variable x over $\mathbb{C}$, let $p_k(H)$ be the greatest common divisor, with leading coefficient 1, of all the $k \times k$ minors of H. Let $p_k(H) = 1$ for $k \leqq 0$. In addition, if q is any polynomial in x, let δq be the degree of q.

LEMMA 2.2. *Let M be a $g \times g$ nilpotent matrix, with invariants n_k, $k \geqq 1$. Then, for $1 \leqq k$, $x^{n_k} = p_{g-k+1}(xI - M)/p_{g-k}(xI - M)$, where I is the $g \times g$ identity matrix.*

This lemma follows from a direct application of the theory of the rational canonical form, as developed in, for example, [6, Chapter VI], to a nilpotent matrix. It follows immediately from this lemma that each $p_k(xI - M)$ is some power of x. Since $p_g(xI - M) = x^g$, we obtain the next result by multiplying the first k x^{n_i}'s together.

LEMMA 2.3. *With the same hypotheses as above, for $1 \leqq k$, $p_k(xI - M)$ is a power of x and $\sum_{i=1}^{k} n_i = g - \delta p_{g-k}(xI - M)$.*

Proof of Theorem 2.1. Let g be the number of vertices in Γ. Let $\hat{M} = xI - M_\Gamma$, and let $\delta p_k = \delta p_k(\hat{M})$. The proof of the theorem rests upon the claim that, for $0 \leqq k$, $\delta p_{g-k} = g - \hat{d}_k$. For, combining this with Lemma 2.3, we have

$$\sum_{i=1}^{k} n_i = g - \delta p_{g-k} = \hat{d}_k = \sum_{i=1}^{k} \hat{\Delta}_i$$

for $1 \leqq k$. From this, $n_k = \hat{\Delta}_k$ follows immediately.

As for proving the claim, since $p_{g-k}(\hat{M})$ must be a power of x, we are looking for the largest power of x that divides every $g - k$ minor of $\hat{M}$, which must equal the smallest power of x appearing as a term in some $g - k$ minor. Since M is generic, this smallest power of x will equal the smallest number of x occurring among all choices of $g - k$ nonzero elements of $\hat{M}$ with no two of them in the same row or column.

Suppose it takes at least A paths to partition Γ into disjoint paths. If $A = g$, Γ can contain no edges and the claim is easy to verify. We can therefore assume that $A < g$. The claim is also clear when $k = 0$ so, for the moment, let us also assume that $0 < k \leqq A$.

Let S be a maximum size k-path in Γ, so $|S| = \hat{d}_k$. Since $k \leqq A$, S must require precisely k paths to be partitioned. Fix a partition of S into k paths. If $a, b, c, \cdots, d, e$ is one of these paths on at least two vertices, we associate this path with the nonzero, non-x elements of $\hat{M}$ in positions $(a, b), (b, c), \cdots, (d, e)$. If we do this for each path in the partition of S with at least two vertices, we end up with $\hat{d}_k - k$ such entries of $\hat{M}$. Since the paths are disjoint, no two of these entries lie in the same row or column. Now, add to these the $g - \hat{d}_k$ entries in the positions (j, j), where j is a vertex not in S. This gives us a collection of $g - k$ nonzero entries of $\hat{M}$, no two in the same row or column, among which are $g - \hat{d}_k$ x's. Hence, $\delta p_{g-k} \leqq g - \hat{d}_k$.

Conversely, suppose we are given $g - k$ entries of $\hat{M}$, no two in the same row or column, with $\delta = \delta p_{g-k}$ x's among them. The $g - k - \delta$ off-diagonal entries correspond, reversing the process used above, to, say, l disjoint paths in Γ, each containing at least two vertices. These paths will use $g - k - \delta + l$ vertices of Γ. The δ diagonal entries will correspond to δ other vertices in Γ. Thus, we must have $(g - k - \delta + l) + \delta \leqq g$, or $l \leqq k$. Besides the $g - k - \delta + l$ vertices of Γ in the l paths, there remain $k + \delta - l$ vertices in Γ. Adding any $k - l$ of these, each considered as a path with 1 vertex, to the l paths above, we obtain a k-path of size $g - \delta$. This quantity can be no more than $\hat{d}_k$; hence, $g - \hat{d}_k \leqq \delta$. Thus, for $0 \leqq k \leqq A$, we have $g - \hat{d}_k = \delta p_{g-k}$.

From this, we see that $\delta p_{g-A} = g - \hat{d}_A = 0$ and $p_{g-A}(\hat{M}) = 1$. By Lemma 2.2, $p_{g-k}(\hat{M})$ divides $p_{g-A}(\hat{M})$ for $k > A$. Thus $\delta p_{g-k} = 0$ for $k > A$. In addition, since $\hat{d}_A = g$, $\hat{d}_k = g$ for $k > A$. Hence, $\delta p_{g-k} = g - \hat{d}_k$ for all $k > A$ as well, completing the proof of our claim and the theorem. □

We note that, as might be expected, any two generic matrices of Γ are similar. As a somewhat unexpected result, we have the following corollary.

COROLLARY 2.4. (cf. [9]). *For all* $k \geqq 1$, $\hat{\Delta}_k \geqq \hat{\Delta}_{k+1}$.

At this point, we would like to introduce some definitions from the theory of partitions (cf. [2]). A *partition* λ is an infinite, nonincreasing sequence of nonnegative integers $\lambda_1 \geqq \lambda_2 \geqq \lambda_3 \geqq \cdots$ with at most finitely many nonzero terms. The λ_i's are the *parts* of the partition, and λ is said to be a *partition of* n, where $n = \sum_{i=1}^{\infty} \lambda_i$. The *Ferrers graph* $F(\lambda)$ of λ is the set of all ordered pairs (i, j) such that $1 \leqq j \leqq \lambda_i$. The *conjugate partition* is the sequence $\sigma_1 \geqq \sigma_2 \geqq \sigma_3 \geqq \cdots$, where σ_j equals the number of λ_i greater than or equal to j.

Thus, Corollary 2.4 asserts that the sequence $\hat{\Delta}$ is a partition of g. Let $\Delta' = \Delta'(\Gamma)$ denote the sequence $\Delta'_1 \geqq \Delta'_2 \geqq \Delta'_3 \geqq \cdots$ that is the conjugate partition of $\hat{\Delta}$, and let $d'_k = d'_k(\Gamma) = \sum_{i=1}^{k} \Delta'_1$ for $k \geqq 1$. At present, we can give an algebraic interpretation to d'_k. A combinatorial interpretation has been given by Saks in [21]. Later, in § 6, we shall mention the special case of this interpretation that holds when Γ is transitive.

COROLLARY 2.5. *The nullity of* M_Γ^k *equals* $d'_k(\Gamma)$ *for* $k \geqq 1$.

Proof. It can be easily shown (cf. [20]) that, if M is nilpotent with invariants n_k, $k \geqq 1$, the nullity of M^k equals $\sum_{i=1}^{\infty} \min\{k, n_i\}$. This expression is the same as the sum of the first k parts of the conjugate partition of $n_1, n_2, n_3, \cdots$. Adding these remarks to Theorem 2.1 and the definitions given above, we arrive at the corollary. □

3. The Young tableau of an A-digraph. Let λ be a partition of n. A (*standard Young*) *tableau of shape* λ is an array (m_{ij}), indexed by all the pairs (i, j) in $F(\lambda)$, whose n entries consist of every integer from 1 to n, and such that $m_{ij} < m_{ij+1}$ and $m_{ij} < m_{i+1j}$. Tableaux play a significant role in several areas of combinatorics [3], [7], [14], [22], [23] and in the representation theory of groups [15], [16].

In the preceding section, we have seen that an A-digraph on g vertices can be associated with a partition $\hat{\Delta}$ of g. For certain labeled A-digraphs, arising from

permutations (cf. § 6), it is known that one can, in fact, associate the graph with a tableau of shape $\hat{\Delta}$, using a construction due to Schensted [22]. We now want to show that one can associate a tableau of shape $\hat{\Delta}(\Gamma)$ with any naturally labeled A-digraph Γ. (An A-digraph is *naturally* labeled if $i<j$ whenever (i, j) is an edge). We begin with an algebraic lemma.

LEMMA 3.1. (cf. [16, p. 91]). *Let* $\mathbf{V}$ *be a finite dimensional vector space. Let* $\mathbf{T}$: $\mathbf{V}\to\mathbf{V}$ *be a nilpotent linear map on* $\mathbf{V}$ *with invariants* n_k, $k\geqq 1$. *Let* $\mathbf{W}$ *be an invariant subspace of* $\mathbf{V}$, *so that* $\mathbf{TW}\subseteq\mathbf{W}$. *We can then view* $\mathbf{T}$ *as a nilpotent map* $\mathbf{T}$: $\mathbf{V/W}\to\mathbf{V/W}$ *with invariants* m_k, $k\geqq 1$. *Then* $n_k\geqq m_k$ *for all* k.

Proof. Let n_k^* and m_k^*, $k\geqq 1$, be the respective conjugates of n_k and m_k, $k\geqq 1$, viewed as partitions. It then suffices to show that $n_k^*\geqq m_k^*$ for all k. As noted in the proof of Corollary 2.5, the nullity of $\mathbf{T}^k$ equals $\sum_{i=1}^{k} n_i^*$. Hence, $n_k^*=$ nullity $(\mathbf{T}^k)-$nullity $(\mathbf{T}^{k-1})=(\dim\mathbf{V}-\dim\mathbf{T}^k\mathbf{V})-(\dim\mathbf{V}-\dim\mathbf{T}^{k-1}\mathbf{V})=\dim(\mathbf{T}^{k-1}\mathbf{V}/\mathbf{T}^k\mathbf{V})$. Similarly, $m_k^*=\dim(\mathbf{T}^{k-1}(\mathbf{V/W})/\mathbf{T}^k(\mathbf{V/W}))$.

To complete the proof, we note that we have the canonical onto maps $\mathbf{T}^{k-1}\mathbf{V}\to\mathbf{T}^{k-1}(\mathbf{V/W})$ and $\mathbf{T}^{k-1}(\mathbf{V/W})\to\mathbf{T}^{k-1}(\mathbf{V/W})/\mathbf{T}^k(\mathbf{V/W})$. These combine to give the onto map ϕ: $\mathbf{T}^{k-1}\mathbf{V}\to\mathbf{T}^{k-1}(\mathbf{V/W})/\mathbf{T}^k(\mathbf{V/W})$. In addition, $\mathbf{T}^k\mathbf{V}$ is contained in the kernel of ϕ. This induces an onto map of vector spaces

$$\mathbf{T}^{k-1}\mathbf{V}/\mathbf{T}^k\mathbf{V}\to\mathbf{T}^{k-1}\mathbf{V}/\text{kernel }\phi\cong\mathbf{T}^{k-1}(\mathbf{V/W})/\mathbf{T}^k(\mathbf{V/W}).$$

Since the map is onto, we must have $n_k^*\geqq m_k^*$, as required. □

THEOREM 3.2. Let Γ be an A-digraph, and let x be a source or sink in Γ. Let Γ' be Γ with x deleted. Then $\hat{\Delta}_k(\Gamma)\geqq\hat{\Delta}_k(\Gamma')$ for all k.

Proof. Let M be a generic matrix for Γ. Let M' be M with the row and column corresponding to x removed. Then M' is a generic matrix for Γ'. Let $\mathbf{V}$ be the vector space over $\mathbb{C}$ generated freely by the vertices of Γ, and let $\mathbf{T}$ be the nilpotent linear map on $\mathbf{V}$ corresponding to the matrix M. Define $\mathbf{V}'$ and $\mathbf{T}'$ similarly, using Γ' and M'.

For the moment, assume that x is a sink. Then $\mathbf{T}x=0$ and the subspace $\langle x\rangle$, generated by x, is invariant. It is easy to see that $\mathbf{V}/\langle x\rangle$ is isomorphic to $\mathbf{V}'$, and that $\mathbf{T}$ acts on $\mathbf{V}/\langle x\rangle$ in the same manner as $\mathbf{T}'$ on $\mathbf{V}'$. Hence, the invariants of $\mathbf{T}'$ are the same as the invariants of $\mathbf{T}$ acting on $\mathbf{V}/\langle x\rangle$. If we now invoke Theorem 2.1 and Lemma 3.1, we obtain the desired result.

If x is a source, reverse the edges in Γ and Γ'. This does not affect $\hat{\Delta}(\Gamma)$ or $\hat{\Delta}(\Gamma')$, but x is now a sink and we can apply the previous case. □

This theorem was originally given in terms of posets by Fomin [5], though the proof presented here is entirely different from his, which used network flows.

COROLLARY 3.3. *If a source or a sink occurs in every maximum-sized* k*-path, it occurs in every maximum-sized* h*-path for all* $h\geqq k$.

Proof. Using the notation of Theorem 3.2, since x occurs in every maximum-sized k-path, we must have $\hat{d}_k(\Gamma)>\hat{d}_k(\Gamma')$. By the theorem, $\hat{\Delta}_j(\Gamma)\geqq\hat{\Delta}_j(\Gamma')$ for all j. Thus, $\hat{d}_h(\Gamma)=\hat{d}_k(\Gamma)+\sum_{j=k+1}^{h}\hat{\Delta}_j(\Gamma)>\hat{d}_k(\Gamma')+\sum_{j=k+1}^{h}\hat{\Delta}_j(\Gamma')=\hat{d}_h(\Gamma')$. This strict inequality forces x to be in every maximum-sized h-path. □

As suggested by Fomin, we can use Theorem 3.2 to construct a tableau from an A-digraph. Let Γ be a naturally labeled A-digraph on g vertices. Let Γ' be Γ with vertex g removed. Assume we have constructed a tableau $Y(\Gamma')$ of shape $\hat{\Delta}(\Gamma')$. By Theorem 3.2, $F(\hat{\Delta}(\Gamma))$ contains $F(\hat{\Delta}(\Gamma'))$ plus one more pair, say, (i, j). Define $Y(\Gamma)$ to be $Y(\Gamma')$ with g adjoined in position (i, j). $Y(\Gamma)$ will have shape $\hat{\Delta}(\Gamma)$ by construction.

In § 6, we shall show that this construction extends the Schensted construction mentioned above.

How much information about Γ is contained in $Y(\Gamma)$? Certainly, many different labeled graphs are sent to the same tableau. For each natural labeling of Γ, we obtain a (possibly new) tableau. However, the set of tableaux obtained in this manner still does not determine the graph. The graphs shown in Fig. 1 both have the same associated set of tableaux. If, instead, we consider the multiset of such tableaux, we can distinguish the graphs. Does this hold in general? Given the multiset (or a special subcollection) of tableaux associated with the natural labelings of an A-digraph, does this distinguish the graph from any other or, even better, can we easily reconstruct the graph from the tableaux?

FIG. 1

At present, we can only offer a small step in answering these questions. If n occurs in position (i, j) of $Y(\Gamma)$ and $n-1$ occurs in position (i', j'), we must have either $i' \geqq i$ and $j' < j$, or $i' < i$ and $j' \geqq j$. If $(n-1, n)$ is not an edge in Γ, we cannot tell which possibility occurs. For example, consider the graphs and their associated tableaux shown in Fig. 2.

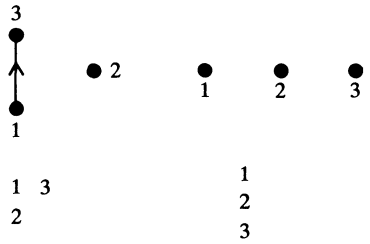


FIG. 2

However, we do have the following.

THEOREM 3.4. *Using the above notation, if* $(n-1, n)$ *is an edge in* Γ, *we must have* $i' \geqq i$ *and* $j' < j$.

Proof. By the construction of $Y(\Gamma)$, we can assume that Γ has n vertices. Let Γ' be Γ with vertex n deleted. Assume that $i' < i$. Then $\hat{d}_{i'}(\Gamma) = \hat{d}_{i'}(\Gamma')$. Now, $n-1$ occurs in every maximum-sized i'-path in Γ'. Let S be such an i'-path, so $|S| = \hat{d}_{i'}(\Gamma')$. Then $S \cup \{n\}$ is an i'-path in Γ, since Γ is naturally labeled and $(n-1, n)$ is an edge in Γ. This implies $\hat{d}_{i'}(\Gamma) \geqq |S \cup \{n\}| = |S| + 1 = \hat{d}_{i'}(\Gamma') + 1 = \hat{d}_{i'}(\Gamma) + 1$, a contradiction. Hence, $i' \geqq i$, forcing $j' < j$. □

4. On *k*-saturated partitions. For a given naturally labeled A-digraph Γ, it is possible to give an alternate construction for $Y(\Gamma)$. Before we can get to this consruction, we need a result that is significant in its own right and that generalizes a known result concerning posets. As usual, we begin with some algebra.

If $A = (a_{ij})$ and $B = (b_{ij})$ are two matrices of the same size, we order them by letting $A < B$ if $b_{ij} = 0$ implies that $a_{ij} = 0$.

LEMMA 4.1. *Let A and B be two $n \times n$ matrices with complex entries such that $A < B$ and B is generic. Then nullity $A^k \geqq$ nullity B^k for all $k \geqq 1$.*

Proof. Let B' be a singular submatrix of B^k. Since B is generic and $\det B'$ is a polynomial in the nonzero entries of B with rational coefficients, the polynomial must be identically zero. This forces $\det A' = 0$, where A' is the submatrix of A^k corresponding to B' in B^k. Hence, a nonsingular submatrix of A^k corresponds to a nonsingular submatrix of B^k, and rank $A^k \leqq$ rank B^k. Equivalently, nullity $B^k \leqq$ nullity A^k.

Let Γ be an A-digraph on g vertices, and let $\mathcal{P}$ be a partition of Γ into paths. Let $B_k(\mathcal{P}) = \sum \min\{k, |P|\}$, the sum being taken over all paths P in $\mathcal{P}$. Using $\mathcal{P}$, we can define the following $g \times g$ matrix. Let $M = (m_{ij})$ with $m_{ij} = 1$ if (i, j) is an edge occurring in some path in $\mathcal{P}$ and $m_{ij} = 0$ otherwise. Clearly, $M < M_\Gamma$ and hence, $d'_k(\Gamma) =$ nullity $M_\Gamma^k \leqq$ nullity M^k. In addition, it is easy to see that the nullity of M^k equals $B_k(\mathcal{P})$. Thus, $B_k(\mathcal{P})$ provides an upper bound for $d'_k(\Gamma)$. We call $\mathcal{P}$ *k-saturated* if $d'_k(\Gamma) = B_k(\mathcal{P})$.

The main result of this section is that, for any given k, there exists a k-saturated partition. To demonstrate this, we will mimic proofs used by Saks ([19] and [21, pp. 78–81]) to prove related results.

Let $e_k(\Gamma) = \min\{B_k(\mathcal{P})\}$, the minimum taken over all partitions $\mathcal{P}$ of Γ into paths. We want to show that $e_k(\Gamma) = d'_k(\Gamma)$. This definitely holds if $k = 1$, for then both sides count the smallest number of paths necessary to partition Γ. As noted above, we also have $d'_k(\Gamma) \leqq e_k(\Gamma)$.

For the next piece of the proof, we need the following lemma.

LEMMA 4.2. *Let M be a nilpotent $n \times n$ matrix. For $k \geqq 1$, let M_k be the matrix of order kn with k copies of M down the main diagonal, $k-1$ copies of the order n identity matrix I down the superdiagonal and 0's elsewhere. Then nullity $M_k =$ nullity M^k.*

Proof. We first assume that M is an $n \times n$ Jordan block, with $n-1$ 1's down the superdiagonal and 0's elsewhere. To compute the rank of M_k, we can use elementary row and column operations. First, use the $n-1$ 1's in the first M block to eliminate the first $n-1$ 1's in the first I block. The remaining 1 in this I block can be used to eliminate the last 1 in the second M block. The remaining $n-2$ 1's in this block can then be used to eliminate the first $n-2$ 1's in the second I block. The remaining two 1's in this block can then be used to eliminate the last two 1's from the third M block. If $k \leqq n$, this alternating elimination process ends at the kth M block, with $n-k$ 1's remaining in this block. At this stage, all the 1's lie on different rows and columns, and a simple count shows that there are $k(n-1)$ of these 1's. Hence, if $k \leqq n$, the rank of M_k equals $k(n-1)$.

If $k > n$, this alternating elimination process ends at the nth M block, which will have all of its 1's eliminated. At this point, we can use the 1's in the nth I block to eliminate all the 1's in the $(n+1)$th M block. Then we can use the 1's in the $(n+1)$th I block to eliminate all the 1's in the $(n+2)$th M block. This process can be continued until all the 1's are eliminated from the kth M block. All the 1's in this final matrix lie on different rows and columns. The first n^2 columns contribute $n(n-1)$ 1's, while each successive n columns contributes n more 1's. Thus, we end up with $n(n-1) + (k-n)n = n(k+1)$ 1's. Hence, if $k \geqq n$, the rank of M_k equals $n(k-1)$.

Now it is easy to compute that the rank of M^k is $n-k$ if $k \leqq n$ and is 0 if $k \geqq n$. These results imply that $n(k-1) +$ rank $M^k =$ rank M_k for all k. Hence, nullity $M_k = nk -$ rank $M_k = n -$ rank $M^k =$ nullity M^k, and the lemma holds if M is a Jordan block.

Next, let A and B be two nilpotent matrices, and assume the lemma holds for them. Let $C = A \oplus B$, the direct sum of A and B. Then C is nilpotent and $C^k = A^k \oplus B^k$. Hence, nullity $C^k =$ nullity $A^k +$ nullity B^k. On the other hand, it is easy to see

that C_k is similar, via a permutation matrix, to $A_k \oplus B_k$. This implies that the nullity of C_k equals nullity $(A_k \oplus B_k)$ = nullity A_k + nullity B_k. Since the lemma holds for A and B, we see that it also holds for C. Combining this result with the preceding result, we find that the lemma holds if M is in Jordan canonical form.

Finally, given any nilpotent M, let $N = A^{-1}MA$ be its Jordan canonical form. Let B be the direct sum of k copies of A, so that B^{-1} is the direct sum of k copies of A^{-1}. It is easy to verify that $B^{-1}M_kB = N_k$. Hence, we have nullity M_k = nullity $(B^{-1}M_kB)$ = nullity N_k. But this we know to be nullity N^k = nullity $(A^{-1}MA)^k$ = nullity $A^{-1}M^kA$ = nullity M^k. Thus, the lemma holds for all nilpotent M.

To apply this lemma, given an A-digraph Γ on g vertices, let Γ_k be the A-digraph on the set of vertices (x, i), where x is a vertex of Γ and $1 \leqq i \leqq k$. The only edges in Γ_k go from (x, i) to $(x, i+1)$, or from (x, i) to (y, i), where (x, y) is an edge in Γ. If x is labeled j in Γ, label (x, i) by $j + g(i-1)$ in Γ_k.

Now, let M be a generic matrix of Γ, and let N be a generic matrix for Γ_k. Then $M_k < N$. Using the previous two lemmas, we obtain nullity $N \leqq$ nullity M_k = nullity M^k. This gives us the next piece in our proof.

LEMMA 4.3. $d'_1(\Gamma_k) \leqq d'_k(\Gamma)$.

Next, we need the following result of Saks, whose proof can be found in [21, pp. 78–81].

LEMMA 4.4. $e_k(\Gamma) = e_1(\Gamma_k)$.

To put the pieces of the proof together, we recall Lemma 4.3, which, along with some earlier remarks, gives us $e_1(\Gamma_k) = d'_1(\Gamma_k) \leqq d'_k(\Gamma) \leqq e_k(\Gamma)$. Adding this to Lemma 4.4, we obtain the desired result. □

THEOREM 4.5. *For all k,* $d'_k(\Gamma) = e_k(\Gamma)$.

COROLLARY 4.6. $d'_1(\Gamma_k) = d'_k(\Gamma)$.

Another proof of Theorem 4.5 and its corollary is given in [21, Chapter VI].

5. A characterization of $Y(\Gamma)$. We can now proceed apace to the promised variant construction of $Y(\Gamma)$ for a given naturally labeled A-digraph Γ on g vertices. We have already seen that $g - d'_k$ is the rank of M^k_Γ. This new construction will be based on another interpretation of $g - d'_k$.

A *k-matching* in Γ is a k-columned matrix $X = (x_{ij})$ with, say, m rows, in which the x_{ij} are labels of the vertices of Γ, (x_{ij}, x_{ij+1}) is an edge in Γ for all i and j, and the entries in a given column are distinct. The set $\{x_{1k}, x_{2k}, \cdots, x_{mk}\}$ we call a *k-source* and the *source* of the given k-matching. The terminology here is due to Greene [11].

Let $\mathcal{P}$ be a partition of Γ such that $B_k(\mathcal{P}) = d'_k(\Gamma)$. Let $x_1, x_2, \cdots, x_l$ be a path in $\mathcal{P}$ with $l \geqq k+1$. Then the sequences $x_1, x_2, \cdots, x_{k+1}$; $x_2, x_3, \cdots, x_{k+2}$; $\cdots$; x_{l-k}, $x_{l-k+1}, \cdots, x_l$ form the rows of a $(k+1)$-matching. If we do this for each such path in $\mathcal{P}$ and combine all the rows together, we obtain a $(k+1)$-matching with $g - B_k(\mathcal{P}) = g - d'_k$ rows, as first noted by Greene. Hence, the size of the largest source of a $(k+1)$-matching in Γ is at least $g - d'_k$. We shall now show that these two quantities are equal.

THEOREM 5.1. *Let Γ be an A-digraph on g vertices. The largest size of a $(k+1)$-source in Γ equals $g - d'_k$.*

Proof. Let $X = (x_{ij})$ be an m-rowed $(k+1)$-matching in Γ. Let N be a generic matrix for Γ_k. N contains k $g \times g$ blocks along the main diagonal, each a generic matrix for Γ. Call these blocks $C_1, C_2, \cdots, C_k$, ordering them from bottom to top. (Thus, C_1 is the bottom rightmost block.) In addition, N contains $k-1$ $g \times g$ blocks along the superdiagonal, each a nonsingular diagonal generic matrix. Call these blocks $I_1, I_2, \cdots, I_{k-1}$, again starting at the bottom right, and labeling up and to the left.

Consider the following nonzero entries in N. For all i and j with $1 \leqq i \leqq m$ and $1 \leqq j \leqq k$, take the (x_{ij}, x_{ij+1}) entry in C_j. In addition, for $1 \leqq j < k$, take all the (z, z) entries in I_j for which z does not appear in column $j+1$ of X. We leave to the reader the easy verification that no two of these entries lie in the same row or column. We have $mk + (g-m)(k-1) = m + g(k-1)$ chosen entries and, since N is generic, the rank of N is at least $m + g(k-1)$. Hence, $m \leqq g - \text{nullity } N = g - d'_1(\Gamma_k) = g - d'_k(\Gamma)$, by Corollaries 2.5 and 4.6. Since we have already seen that $m \geqq g - d'_k$, the proof is complete. □

The restriction of this theorem to posets was first proved by Greene in [10]. The theorem says that the largest size of a source of a k-matching in Γ equals the number of elements in $F(\Delta(\Gamma))$ in columns k through $\hat{\Delta}_1(\Gamma)$, or, if Γ is naturally labeled, the number of entries in $Y(\Gamma)$ in the same columns. Greene conjectured that, in fact, the entries in the last columns of $Y(\Gamma)$ form a special, maximum-sized source, and proved this for the graphs covered by the Schensted construction. We will now verify his conjecture for general A-digraphs.

We order the collection of all maximum-sized k-sources in Γ lexicographically, that is, if A and B are such k-sources, we have $A < B$ if and only if the least element that occurs in only one of A and B appears in A. Thus, the collection is a chain and possesses a unique minimum.

THEOREM 5.2. *Let Γ be a naturally labeled A-digraph. For $1 \leqq k \leqq \hat{\Delta}_1(\Gamma)$, the union of the entries in columns k through $\hat{\Delta}_1$ of $Y(\Gamma)$ equals the minimum maximum-sized k-source in Γ.*

Proof. For any naturally labeled A-digraph ψ, let $C_k(\psi)$ be the entries in columns k through $\hat{\Delta}_1(\psi)$ of $Y(\psi)$.

Let Γ contain g vertices. We use induction on g, the case $g = 1$ being clear. Assuming $g > 1$, let Γ' be Γ with vertex g deleted. By induction, $C_k(\Gamma')$ equals the minimum maximum-sized k-source in Γ'. In addition, if g occurs in position (i, j) of $Y(\Gamma)$, $Y(\Gamma)$ is formed by adjoining g to $Y(\Gamma')$ in position (i, j).

It will be useful to note that the k-sources of Γ form the independent sets of a matroid, and thus, any k-source can be extended to a maximum-sized one. To see this, define a k-partite undirected graph on the vertex sets $V_m = \{1_m, 2_m, \cdots, g_m\}$ for $1 \leqq m \leqq k$, with an edge joining r_l to s_m if and only if $m = l+1$ and (r, s) is an edge in Γ. Then a k-matching in Γ, whose last column consists of the vertices $\{x, y, \cdots, z\}$, corresponds to a matching of $\{x_k, y_k, \cdots, z_k\}$ with certain vertices in V_{k-1}, which are in turn matched with certain vertices in V_{k-2}, etc. It is well known [1, pp. 276–281] that all such sets $\{x_k, y_k, \cdots, z_k\}$ form the independent sets of a matroid, generalizing the transversal matroids.

Now, if $j < k$, we have $C_k(\Gamma) = C_k(\Gamma')$. In addition, Theorem 5.1 tells us that the maximum size of a k-source in Γ equals the maximum size of one in Γ'. Let $\{g, x, y, \cdots, z\}$ be a maximum-sized k-source in Γ containing g. Then $\{x, y, \cdots, z\}$ is a k-source in Γ' and can be extended to a maximum-sized one $\{w, x, y, \cdots, z\}$ in Γ' and hence, in Γ. We thus have $\{w, x, y, \cdots, z\} < \{g, x, y, \cdots, z\}$, and the former does not contain g. From this, we see that the minimum maximum-sized k-source of Γ equals the minimum one in Γ', which equals $C_k(\Gamma') = C_k(\Gamma)$, as required.

Finally, for $j \geqq k$, $C_k(\Gamma) = C_k(\Gamma') \cup \{g\}$, and, from Theorem 5.1, the maximum size of a k-source in Γ is one more than the maximum size of one in Γ'. Because of this difference, every maximum-sized k-source in Γ must contain g. Then, if $\{g, x, y, \cdots, z\}$ is a maximum-sized k-source in Γ, $\{x, y, \cdots, z\}$ is one in Γ'. Conversely, if $\{x, y, \cdots, z\}$ is a maximum-sized k-source in Γ', it is a k-source in Γ and can be extended to a maximum-sized one in Γ, which must necessarily be $\{g, x, y, \cdots, z\}$. Hence, the

minimum maximum-sized k-source in Γ equals the minimum such source for Γ' with g added. By induction, this is $C_k(\Gamma') \cup \{g\} = C_k(\Gamma)$, completing the proof. □

As noted by Greene, since sources are independent sets in a matroid, to find the minimum maximum-sized k-source, one can use a greedy algorithm, picking the least vertex that is a source, then continuing to add the least vertex that allows the set to remain a k-source.

6. Transitive A-digraphs. In this and the next section, we consider some interesting results that occur if we assume that Γ is transitive, that is, a poset. What we have been calling paths in a digraph become chains in a poset. In addition, we have the important notion of an *antichain* as a set of vertices all of whose elements are unrelated in the poset.

The definitions of $\hat{d}_k$, $\hat{\Delta}_k$, d'_k, Δ'_k remain the same. However, we now have a combinatorial interpretation for d'_k as well as $\hat{d}_k$. For $k \geqq 1$, we define a *k-family* in a poset Γ as a set of vertices containing no chain of size $k+1$. Thus, a 1-family is the same as an antichain. The subject of k-families has been extensively researched by Greene and Kleitman [8], Greene [9] and others. We shall be content with noting a couple of known results as they occur in the present context.

Let $d_k = d_k(\Gamma)$ be the maximum size of a k-family in Γ for $k \geqq 1$ and let $d_0 = 0$. Let $\Delta_k = d_k - d_{k-1}$ for all $k \geqq 1$, and let Δ represent the sequence $\Delta_1, \Delta_2, \Delta_3, \cdots$ of nonnegative integers. We saw in § 4 that $e_k(\Gamma) = d'_k(\Gamma)$. But, for posets, it is well known (e.g. [8], [19]) that $e_k(\Gamma)$ also equals $d_k(\Gamma)$. This leads us to the following results. Theorem 6.1 was essentially proved by Saks in [20], (cf. also [21]).

THEOREM 6.1. *Given a poset* Γ, *we have* $d'_k(\Gamma) = d_k(\Gamma)$ *for all* k.

COROLLARY 6.2. (Greene and Kleitman [8]). *For all* $k \geqq 1$, $\Delta_k \geqq \Delta_{k+1}$.

COROLLARY 6.3. (Greene [9]). *The partition* Δ *is the conjugate partition of* $\hat{\Delta}$.

Thus, for a poset Γ, we have two combinatorial methods for arriving at the same partition. This would seem to indicate that the partition $\hat{\Delta}$ and, if Γ is naturally labeled, the tableau $Y(\Gamma)$ are natural objects to associate with the poset. This relation is even stronger when we consider certain special posets with certain special labelings. Let π be a permutation of the set $\{1, 2, \cdots, g\}$. Using this permutation, we can construct two naturally labeled posets. Define Γ_1 by letting vertex i be less than vertex j in Γ_1 if $i < j$ and $\pi(i) < \pi(j)$. Define Γ_2 similarly, using π^{-1} instead of π. We then have two tableaux $Y(\Gamma_1)$ and $Y(\Gamma_2)$ of the same shape.

Using an entirely different approach, based on an easy, efficient algorithm, Schensted [22] gave a one-to-one correspondence associating π with a pair of tableaux $P(\pi)$ and $Q(\pi)$ of the same shape. For the interested reader, there is an extensive literature on this correspondence, its properties and its applications. (See [7], [23], and the references cited there). Here, we simply remark that $P(\pi) = Y(\Gamma_1)$ and $Q(\pi) = Y(\Gamma_2)$. This can most easily be seen by noting that Greene (Theorem 3 and the following remarks in [11]) has characterized P and Q in precisely the same manner in which we characterized Y in Theorem 5.2.

When Γ is derived, as above, from a permutation, it is a two-dimensional poset [4] with a special labeling. It is not surprising, therefore, to find properties of $Y(\Gamma)$ that hold in this case, but not for arbitrary posets, or arbitrary natural labelings. With this in mind, one can ask, with little optimism, for a way to extend the sleek Schensted construction of P and Q to arbitrary posets, thereby replacing the rather tedious ways of constructing $Y(\Gamma)$ given above.

7. Posets and k-scatters. In Theorems 5.1 and 5.2, we were able to determine the number of entries and the entries themselves in the last columns of $Y(\Gamma)$. We

now want to consider whether analogous results hold, when Γ is a poset, for the last rows of $Y(\Gamma)$. As shown by Greene [11], such analogues do exist when Γ and its labeling are derived from a permutation. For general posets, we must answer with a yes and a no.

Let Γ be a naturally labeled poset. A *k-scatter* is a k-columned matrix $X=(x_{ij})$ in which the x_{ij} are labels of the vertices of Γ; for all i and j, $x_{ij}<x_{ij+1}$ as integers; the entries in any given column of X are distinct; and there exists no sequence x_{ar}, $x_{as}=x_{bt}$, $x_{bu}=x_{cv}$, $\cdots$, $x_{dw}=x_{ey}$, x_{ez} such that $r\leqq s$, $t\leqq u$, $\cdots$, $y\leqq z$ and vertex x_{ar} is less than vertex x_{ez} in Γ. In particular, the last condition implies that the vertices corresponding to a row in a k-scatter form an antichain.

THEOREM 7.1. *Let Γ be a naturally labeled poset on g vertices. The largest number of rows occurring in a $(k+1)$-scatter is $g-\hat{d}_k$.*

Half of the proof of this theorem requires an easy imitation of half of the proof of Theorem 5.1. However, the other half requires some combinatorial groundwork, to which we proceed and postpone the theorem's proof until later.

Let m and k be positive integers, and let T be a finite subset of the positive integers. Consider an $m\times k$ matrix $A=(a_{ij})$ with the following properties:

(a) The entries of A come from $T\cup\{0\}$.

(b) For $j\geqq 1$, j appears at most once in any row or column.

(c) In each row, the nonzero entries appear in increasing order.

A *segment* in A is a subsequence of a row in A. A *string* in A is a sequence $S_1, S_2, \cdots, S_r$ of disjoint segments such that the last entry in S_i is nonzero and equals the first entry in S_{i+1} for $1\leqq i\leqq r-1$. S_r is the *final segment* of the string. The *weight* of a string is the number of 0's occurring in it. A row is *long* if its rightmost entry occurs in the final segment of some string of weight $\geqq k$.

LEMMA 7.2. *If $|T|\leqq m$, A contains at least $m-|T|$ long rows.*

Proof. The lemma clearly holds if $|T|=0$. We now use induction and assume that the lemma holds if $0\leqq|T|\leqq n<m$. Let $|T|=n+1$. Without loss of generality, we can assume that $T=\{1,2,\cdots,n+1\}$. In addition, we can assume that, if A contains $r\,n+1$'s, they occur in the first r rows and, for $i<j$, the $n+1$ in row i occurs to the left of the $n+1$ in row j.

Define $A'=(a'_{ij})$ by replacing the $n+1$'s in A by 0's. By induction, A' contains at least $m-n$ long rows. If A contains no $n+1$'s, $A=A'$ and we are done since $m-n>m-(n+1)$. Otherwise, insert $n+1$ into the first row of A' in the same position it occupies in A. This might prevent row 1 from being long, but it cannot affect the status of any other row. Hence, we have at least $m-n-1$ long rows, ignoring the first row.

Assume we have inserted $n+1$'s into the first s rows of A' in the same positions they occupy in A, and that we have at least $m-n-1$ long rows, ignoring row s. Now, insert $n+1$ into row $s+1$ to make the row agree with row $s+1$ in A. If row $s+1$ was not long, we still have $m-n-1$ long rows, now ignoring row $s+1$. If row $s+1$ was a long row, after inserting $n+1$, it may or may not be.

However, row s definitely is. To see this, let a'_{s+1l} be the first entry in the last segment of a string of weight $\geqq k$ that exists before $n+1$ is inserted into row $s+1$. We can assume that the final segment is $a'_{s+1l}, a'_{s+1l+1}, \cdots, a'_{s+1k}$. If $n+1$ is inserted in row $s+1$ in position j, and $n+1$ occurs in position i in row s, replace this final segment by the two segments $a_{s+1l}, \cdots, a_{s+1j}$ and $a_{si}, \cdots, a_{sk}$. This drops the $k-j+1$ 0's $a'_{s+1j}, \cdots, a'_{s+1k}$ and adds the $k-i$ 0's $a_{si+1}, \cdots, a_{sk}$. Since $j>1$, the new string has weight at least k and row s is long, as claimed. As row s was not used in the previous count of long rows, we still have at least $m-n-1$ long rows, now ignoring

row $s+1$. Hence, after we have inserted all of the $n+1$'s and obtained A, we still have $m-n-1$ long rows, as required.

Proof of Theorem 7.1. A theorem of Greene [9, Thm. 1.4] guarantees that we can partition Γ into antichains $A_1, A_2, \cdots, A_r$, for some r, such that $\sum_{i=1}^{r} \min\{|A_i|, k\} = \hat{d}_k$. Now, suppose $|A_i| = l \geqq k+1$, and the vertices of A_i have labels $c_1 < c_2 < \cdots < c_l$. We can then create a $(k+1)$-scatter with rows $c_1, c_2, \cdots, c_{k+1}$; $c_2, c_3, \cdots, c_{k+2}$; $\cdots$; $c_{l-k}, c_{l-k+1}, \cdots, c_l$. If we do this for every A_i, $1 \leqq i \leqq r$, with $|A_i| \geqq k+1$ and put all the rows of the $(k+1)$-scatters together, we obtain a $(k+1)$-scatter with $g - \sum_{i=1}^{r} \min\{|A_i|, k\} = g - \hat{d}_k$ rows. So, the maximum number of rows in a $(k+1)$-scatter is at least $g - \hat{d}_k$.

Let C be a k-path (a collection of at most k disjoint chains) of maximum size in Γ, so that $|C| = \hat{d}_k$. Let $S = \Gamma - C$. We claim that every $(k+1)$-scatter with m rows contains at least m distinct elements of S. This is certainly true if $m = 0$. For $m > 0$, let T be the set of elements in S which occur in a given $(k+1)$-scatter. In the $(k+1)$-scatter, replace the labels of elements in C by 0's. We then have an array of the type discussed above, using $k+1$ instead of k. If $|T| < m$, by Lemma 7.2, the array contains a long row and hence a string of weight $k+1$. Such a string corresponds in the $(k+1)$-scatter to an antichain in Γ with at least $k+1$ elements in C. But C is a k-path and contains no antichain of size $k+1$. This contradiction forces $m \leqq |T|$, as claimed.

From the claim, we have $m \leqq |S| = g - |C| = g - \hat{d}_k$. This, combined with the previously derived inequality, completes the proof. □

It follows from the proof of Theorem 7.1 and the Marriage theorem (cf. [1, Thm. 6.25]) that the rows of a $(k+1)$-scatter can always be matched into the set S as defined in the proof. That is, if a $(k+1)$-scatter X has m rows, we can find m distinct elements $s_1, s_2, \cdots, s_m$ in S such that s_i occurs in row i of X for $1 \leqq i \leqq m$.

Theorem 7.1 provides a row version of Theorem 5.1. To obtain a row version of Theorem 5.2, we could define the source of a k-scatter in the same way in which we defined a source for a k-matching. We can also order the maximum-sized sources as before. It can easily be seen that, if the sources of all k-scatters form the independent sets of a matroid, we can imitate the proof of Theorem 5.2 and show that the entries in rows k through $\Delta_1(\Gamma)$ in $Y(\Gamma)$ make up the minimum maximum-sized k-source.

However, it is possible that the sources are not the independent sets of a matroid. For example, the poset shown in Figure 3 has the 2-sources $\{2, 4, 6\}$ and $\{2, 3, 5, 6\}$, but the former cannot be extended to a 2-source of size 4.

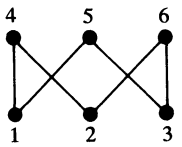


FIG. 3

In the above example, the analogue of Theorem 5.2 holds, despite the failure of the independent set criterion. In general, however, the analogue fails, as with the poset shown in Fig. 4. The entries in rows 4 and 5 of $Y(\Gamma)$ are $\{4, 7\}$, but the only maximum-sized 4-source is $\{6, 7\}$. Empirical evidence suggests that, if a poset is "properly" labeled, the analogue does hold, but we are unable to conjecture precisely

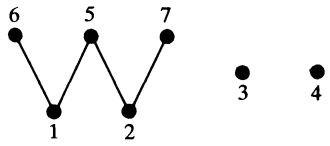


FIG. 4

what "properly" means in this case. Perhaps it would be best to simply look for a "correct" row version of Theorem 5.2.

Acknowledgments. The author would like to express his thanks to Richard Stanley for his suggestions and comments, and to David Gluck for his aid in putting together the algebraic tools.

REFERENCES

[1] M. AIGNER, *Combinatorial Theory*, Springer-Verlag, New York, 1979.
[2] G. ANDREWS, *The Theory of Partitions*, Addison-Wesley, Reading, MA, 1976.
[3] E. A. BENDER AND D. E. KNUTH, *Enumeration of plane partitions*, J. Combin. Theory Ser. A, 13 (1972), pp. 40–54.
[4] B. DUSHNIK AND E. MILLER, *Partially ordered sets*, Amer. J. Math., 63 (1961), pp. 600–610.
[5] S. V. FOMIN, *Finite partially ordered sets and Young tableaux*, Dokl. Akad. Nauk. USSR, 243 (1978), pp. 1144–1147; Soviet Math. Dokl., 19 (1978), pp. 1510–1514.
[6] F. R. GANTMACHER, *The Theory of Matrices*, vol. 1, Chelsea, New York, 1960.
[7] E. R. GANSNER, *Matrix correspondences of plane partitions*, Pacific J. Math., to appear.
[8] C. GREENE AND D. J. KLEITMAN, *The structure of Sperner k-families*, J. Combin. Theory Ser. A, 20 (1976), pp. 41–68.
[9] C. GREENE, *Some partitions associated with a partially ordered set*, J. Combin. Theory Ser. A, 20 (1976), pp. 67–79.
[10] ———, *Sperner families and partitions of a partially ordered set*, Math. Centre (Amsterdam) Tracts, 56 (1974), pp. 91–106.
[11] ———, *Some order-theoretic properties of the Robinson-Schensted correspondence*, in Combinatoire et Representation du Groupe Symetrique, D. Foata, ed., Lecture Notes in Mathematics 579, Springer, New York, 1977, pp. 114–120.
[12] F. HARARY, *Graph Theory*, Addison-Wesley, Reading, MA, 1971.
[13] I. N. HERSTEIN, *Topics in Algebra*, Blaisdell, Waltham, MA, 1964.
[14] D. E. KNUTH, *Permutations, matrices, and generalized Young tableaux*, Pacific J. Math., 34 (1970), pp. 709–727.
[15] D. E. LITTLEWOOD, *The Theory of Group Characters*, 2nd ed., Oxford, Cambridge, 1950.
[16] I. G. MACDONALD, *Symmetric Functions and Hall Polynomials*, Oxford, Cambridge, 1979.
[17] R. PROCTOR, M. SAKS AND D. STURTEVANT, *Product partial orders with the Sperner property*, Discrete Math., 30 (1980), pp. 173–180.
[18] G.-C. ROTA, *On the foundations of combinatorial theory* I. *Theory of Mobius functions*, Z. Wahrsch. Verw. Gebiete, 2 (1964), pp. 340–368.
[19] M. SAKS, *A short proof of the existence of k-saturated partitions for partially ordered sets*, Adv. Math., 33 (1979), pp. 207–211.
[20] ———, *Dilworth numbers, incidence maps and product partial orders*, this Journal, 1 (1980), pp. 211–215.
[21] ———, *Duality properties of finite set systems*, Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1980.
[22] C. SCHENSTED, *Longest increasing and decreasing subsequences*, Canad. J. Math., 13 (1961), pp. 179–191.
[23] R. P. STANLEY, *Theory and applications of plane partitions* I, II, Stud. Appl. Math., 50 (1971), pp. 167–188, 259–279.
[24] ———, *Weyl groups, the hard Lefschetz theorem, and the Sperner property*, this Journal, 1 (1980), pp. 168–184.
[25] O. ZARISKI AND P. SAMUEL, *Commutative Algebra*, vol. 1, Van Nostrand, New York, 1958.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 4, December 1981

0196-5212/81/0204-0010 $01.00/0

EXPECTED NUMBER OF VERTICES OF A RANDOM CONVEX POLYHEDRON*

D. G. KELLY† AND J. W. TOLLE‡

Abstract. Given m points on the unit sphere in n-space, the hyperplanes tangent to the sphere at the given points bound a convex polyhedron with m facets. If the points are chosen independently at random from the uniform distribution on the sphere, the number V_{mn} of the vertices of the polyhedron is a random variable. We obtain an integral expression for EV_{mn} and asymptotic bounds of the form

$$/\alpha^n n^{(n-6)/2} m \leqq EV_{mn} \leqq \beta^n n^{(n-5)/2} m.$$

1. Introduction. This paper deals with random convex polyhedra having m facets in n-dimensional Euclidean space $\mathbb{R}^n$. We define the term "random $m \times n$ polyhedron" (omitting "convex" throughout the paper) by the following chance experiment: choose m points $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_m$ independently from the uniform distribution on the unit sphere S_{n-1} in $\mathbb{R}^n$; at each $\mathbf{p}_i$, construct the hyperplane tangent to S_{n-1} at $\mathbf{p}_i$ and let D_i be the half-space bounded by this hyperplane and containing the origin; let P be the polyhedron $\bigcap_{i=1}^m D_i$. The random polyhedron P has, with probability one, m facets; the number of its vertices is a random variable V_{mn}.

In this paper, we derive the integral expression (3.3) for the expected value EV_{mn} of this random variable, the upper bound (4.1) and consequent asymptotic upper bound (4.2) on this expression, and also the asymptotic lower bounds (4.7) and (4.8). The asymptotic results can be summarized by saying that there exist constants α and β independent of m and n such that, for any sufficiently large fixed value of n, as m increases we have eventually

$$\alpha^n n^{(n-6)/2} m \leqq EV_{mn} \leqq \beta^n n^{(n-5)/2} m.$$

In the concluding section, we give some of the values of EV_{mn} for moderate m and n, computed numerically from (3.3), and we compare these to known upper and lower bounds on the possible number of vertices of an $m \times n$ polytope.

This study is motivated by attempts to investigate average-case behavior of pivoting algorithms for linear programming problems. Because of the difficulty of analyzing directly the number S of pivots needed for a random problem, it seems useful to consider other quantities associated with linear programs and related to S. Quantities that come to mind in this connection are the number V of vertices (basic feasible solutions) and the number F of facets (effective constraints) of the feasible region of a random linear program, which is a random polyhedron. The probability distribution on polyhedra induced by the chance experiment described above guarantees that F will have a given value m; we are then finding a certain conditional expected value of V given $F = m$. It is conditional on other assumptions as well, however, for our random polyhedra are all circumscribed on the unit sphere, and many combinatorial types of polyhedra cannot be realized as polyhedra so circum-

* Received by the editors June 20, 1980, and in revised form April 10, 1981.

† Department of Statistics, Curriculum in Operations Research and Systems Analysis, Department of Mathematics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514. The work of this author was supported by the U.S. Office of Naval Research under contract N00014-76-C-0550.

‡ Curriculum in Operations Research and Systems Analysis, Department of Mathematics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27514. The work of this author was supported in part by the U.S. Office of Naval Research under contract N00014-76-C-0550, and in part by the U.S. Army Research Office under grant DAAG29-79-G-0014.

scribed. Experiments like those in [2] may yet show, however, that polyhedra as generated here represent a statistically large and typical class of polyhedra. In [2], a more general scheme for generating random polyhedra is considered, in which the points $\mathbf{p}_i$ are chosen from various different distributions on R^n rather than the uniform distribution on the unit sphere. (The hyperplane at $\mathbf{p}_i$ is then not tangent to the unit sphere in general, but is the hyperplane through $\mathbf{p}_i$ and orthogonal to the vector from the origin to $\mathbf{p}_i$.) Other authors, notably Renyi and Sulanke [9], W. Schmidt [10], and Mattheiss and B. Schmidt [6], have considered random polyhedra from various viewpoints.

The reader is referred also to the paper of A. Prékopa [8], in which he finds expected numbers of vertices for random polyhedra generated in a manner different from that used here. It will be noticed that the expected values found there are much smaller than those obtained here; indeed, for any fixed value of n they approach zero as m increases. The reason is that Prékopa's method of generating random polyhedra produces the empty polyhedron (which has no vertices) with high probability (approaching 1 as m increases).

2. Formulation. Given an n-vector $\mathbf{a}$ and a real number b, the inequality $\mathbf{a}^T\mathbf{x} \leqq b$ defines a half-space $\mathbb{R}^n$ with bounding hyperplane $H = \{\mathbf{x}: \mathbf{a}^T\mathbf{x} = b\}$. The nonempty intersection of a finite number, say k, of such half-spaces defines a polyhedron in $\mathbb{R}^n$ with at most k facets. A *random polyhedron* is a polyhedron generated by choosing the k n-vectors $\mathbf{a}_1, \cdots, \mathbf{a}_k$ and the k real numbers $b_1, \cdots, b_k$ from some probability distribution. Obviously the structure of random polyhedra will depend on the choice of the probability distribution from which the data are drawn. For examples of certain distribution schemes, the reader is referred to [2], [4].

In this paper, it is desired to consider only random polyhedra in $\mathbb{R}^n$ with a fixed number m of facets. The general scheme given above for generating polyhedra will lead to redundant constraints, especially in the case where k is large relative to n. This difficulty can be avoided if $b_1, \cdots, b_k$ are chosen as appropriate functions of the $\mathbf{a}_1, \cdots, \mathbf{a}_k$. In particular, if each b_j satisfies

$$b_j = |\mathbf{a}_j| = \left(\sum_{i=1}^{n} \mathbf{a}_{ji}^2 \right)^{1/2}, \tag{2.1}$$

then the random polyhedron generated by the vectors $\mathbf{a}_1, \cdots, \mathbf{a}_m$ is circumscribed about the unit sphere S_{n-1} and has exactly m facets. The facet corresponding to $\mathbf{a}_j$ lies in the hyperplane $H_j = \{x: \mathbf{a}_j^T\mathbf{x} = |\mathbf{a}_j|\}$ and is tangent to S_{n-1} at $\mathbf{p}_j = \mathbf{a}_j/|\mathbf{a}_j|$. The polyhedron is thus completely determined by the unit vectors $\mathbf{p}_1, \cdots, \mathbf{p}_m$. Consequently, an alternative generating scheme for this type of random polyhedron is to choose the vectors $\mathbf{p}_1, \cdots, \mathbf{p}_m$ from some distribution on S_{n-1}.

Of special interest here will be the case where $\mathbf{p}_1, \cdots, \mathbf{p}_m$ are independently and uniformly distributed on S_{n-1}. As is well known [3], this choice of the tangent points can be effected by choosing the components of the vectors $\mathbf{a}_j$, $j = 1, \cdots, m$, independently from the standard normal distribution on the real line and the b_j according to (2.1).

Let $\mathbf{p}_1, \cdots, \mathbf{p}_m$ be independent and identically distributed points on S_{n-1}. As discussed above, these points determine a unique polyhedron in $\mathbb{R}^n$, containing S_{n-1} and having exactly m facets. Let $V_{mn}(\mathbf{p}_1, \cdots, \mathbf{p}_m)$ denote the number of vertices of this polyhedron. The function V_{mn} is a random variable on the space of m-tuples of n-vectors, with values in the non-negative integers. The purpose of this paper is to investigate the expected value of V_{mn}, specifically when the $\mathbf{p}_j$ have the uniform distribution on S_{n-1}.

3. Integral expression for EV_{mn}. Fix integers m and n, $2 \leqq n < m$, and let $\mathbf{p}_1, \cdots, \mathbf{p}_m$ be independent random points chosen from some distribution with density $g(\mathbf{p})$ on S_{n-1}. Let H_i be the hyperplane tangent to S_{n-1} at $\mathbf{p}_i$, $i = 1, \cdots, n$, and let $P(\mathbf{p}_1, \cdots, \mathbf{p}_m)$ be the polytope bounded by $H_1, \cdots, H_m$. Then, with probability 1, any n of $H_1, \cdots, H_m$ intersect in a point of $\mathbb{R}^n$; there are $\binom{m}{n}$ such points of intersection, among which are the vertices of $P(\mathbf{p}_1, \cdots, \mathbf{p}_m)$.

Denote by $\mathscr{A}_{mn}$ the family of all n-subsets of $\{1, 2, \cdots, m\}$. For any $A = \{i_1, \cdots, i_n\} \in \mathscr{A}_{mn}$, let V_A be the event that the point of intersection of $H_{i_1}, \cdots, H_{i_n}$ is a vertex of $P(\mathbf{p}_1, \cdots, \mathbf{p}_m)$. Then V_{mn} is the sum of the indicator functions of the events V_A, and so

$$EV_{mn} = \sum_{A \in \mathscr{A}_{mn}} \Pr(V_A).$$

Because the $\mathbf{p}_i$ are identically distributed, we have by symmetry

$$EV_{mn} = \binom{m}{n} \Pr(V_n),$$

where V_n denotes $V_{\{1,\cdots,n\}}$.

Now, with probability 1, $\mathbf{p}_1, \cdots, \mathbf{p}_n$ lie on a unique small hypercircle on S_{n-1}, which divides S_{n-1} into two unequal caps. Let $C(\mathbf{p}_1, \cdots, \mathbf{p}_n)$ be the smaller of these two caps. Then V_n occurs if and only if none of $\mathbf{p}_{n+1}, \cdots, \mathbf{p}_m$ is in $C(\mathbf{p}_1, \cdots, \mathbf{p}_n)$. Given $\mathbf{p}_1, \cdots, \mathbf{p}_n$, therefore, the conditional probability of V_n is

$$\Pr(V_n | \mathbf{p}_1, \cdots, \mathbf{p}_n) = \left(1 - \int_{C(\mathbf{p}_1,\cdots,\mathbf{p}_n)} g(\mathbf{p})\, d\mathbf{p}\right)^{m-n}.$$

To get $\Pr(V_n)$, we multiply by the joint density of $\mathbf{p}_1, \cdots, \mathbf{p}_n$ and integrate; thus

$$EV_{mn} = \binom{m}{n} \int_{S_{m-1}} \cdots \int_{S_{m-1}} h(\mathbf{p}) \cdot d\mathbf{p}_1 \cdots d\mathbf{p}_n, \tag{3.1}$$

where

$$h(\mathbf{p}) = \left(1 - \int_{C(\mathbf{p}_1,\cdots,\mathbf{p}_m)} g(\mathbf{p})\, d\mathbf{p}\right)^{m-n} g(\mathbf{p}_1) \cdots g(\mathbf{p}_n).$$

Now we assume that the common distribution of $\mathbf{p}_1, \cdots, \mathbf{p}_n$ is the uniform distribution on S_{n-1}. In this case,

$$\int_{C(\mathbf{p}_1,\cdots,\mathbf{p}_n)} g(\mathbf{p})\, d\mathbf{p} = \frac{\text{area of } C(\mathbf{p}_1, \cdots, \mathbf{p}_n)}{\text{area of } S_{n-1}} = \frac{I_{n-2}(r)}{2I_{n-2}(\pi/2)},$$

where $r = r(\mathbf{p}_1, \cdots, \mathbf{p}_n)$ is the angular radius of $C(\mathbf{p}_1, \cdots, \mathbf{p}_n)$, $0 < r < \frac{1}{2}\pi$, and where

$$I_k(r) = \int_0^r \sin^k x\, dx, \qquad 0 < r < \tfrac{1}{2}\pi, \quad k = 0, 1, 2, \cdots,$$

so that $2\pi^{(k+1)/2} I_k(r)/\Gamma(\frac{1}{2}(k+1))$ is the area of a cap of angular radius r on S_{k+1}. Thus (3.1) can be rewritten

$$EV_{mn} = \binom{m}{n} E\left(1 - \frac{I_{n-2}(r)}{2I_{n-2}(\pi/2)}\right)^{m-n},$$

where the random variable r is the radius of $C(\mathbf{p}_1, \cdots, \mathbf{p}_n)$.

Now it follows from results of R. A. Miles [7, Thm. 4, p. 368] that $t = \sin^2 r$ has the beta $(\frac{1}{2}(n-1)^2, \frac{1}{2})$ distribution whose density is proportional to $t^{((m-1)2/2)-1}(1-t)^{-1/2}$ on (0, 1). From this, it follows that the density of r is $f_{n(n-2)}(r)$, where

$$f_k(r) = c(k)\sin^k r, \qquad 0 < r < \tfrac{1}{2}\pi,$$

and

$$c(k) = \frac{2\Gamma(\frac{1}{2}k+1)}{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2}k+\frac{1}{2})}, \qquad k = 0, 1, 2, \cdots. \tag{3.2}$$

Notice that

$$c(k) = \begin{cases} \dfrac{2}{\pi} \cdot \dfrac{2 \cdot 4 \cdot \cdots \cdot k}{1 \cdot 3 \cdot \cdots \cdot (k-1)}, & k = 2, 4, 6, \cdots, \\ \dfrac{1 \cdot 3 \cdot \cdots \cdot k}{2 \cdot 4 \cdot \cdots \cdot (k-1)}, & k = 3, 5, 7, \cdots. \end{cases}$$

Moreover, the distribution function of the density f_k is

$$F_k(r) = \frac{I_k(r)}{I_k(\frac{1}{2}\pi)}, \qquad 0 < r < \tfrac{1}{2}\pi.$$

Using the above notation, we obtain the expression

$$EV_{mn} = \binom{m}{n} \int_0^{\pi/2} (1 - \tfrac{1}{2}F_{n-2}(r))^{m-n} f_{n(n-2)}(r)\, dr. \tag{3.3}$$

Although we will not use it, the following alternative expression may be of interest. If B_k and b_k denote the distribution function and density of the beta$(\frac{1}{2}k, \frac{1}{2})$ distribution, then the substitution $t = \sin^2 r$ in (3.3) gives

$$EV_{mn} = \binom{m}{n} \int_0^1 (1 - \tfrac{1}{2}B_{n-1}(t))^{m-n} b_{(n-1)^2}(t)\, dt.$$

4. Upper and lower bounds. The evaluation of (3.3) is not difficult in case n is 2 or 3:

$$EV_{m2} = \binom{m}{2} \int_0^{\pi/2} \left(1 - \frac{r}{\pi}\right)^{m-2} \cdot \frac{2}{\pi}\, dr = m(1 - (\tfrac{1}{2})^{m-1})$$

and

$$\begin{aligned} EV_{m3} &= \binom{m}{3} \int_0^{\pi/2} (1 - \tfrac{1}{2}(1 - \cos r))^{m-3} \cdot \tfrac{3}{2}\sin^3 r\, dr \\ &= 2m - 4 - (\tfrac{1}{2})^{m-1}(m+1)(m-2). \end{aligned}$$

Notice that a *bounded* polyhedron (polytope) with m facets in $\mathbb{R}^2$ has m vertices; in $\mathbb{R}^3$, is has $2m = 4$ vertices (because of Euler's formula and because, with probability 1, each vertex is adjacent to three edges). The above expected values are less than these numbers because of the positive probability that a polyhedron is unbounded, i.e., that $\mathbf{p}_1, \cdots, \mathbf{p}_m$ all lie in one semicircle or hemisphere. Indeed, because m and $m-1$ are the only possible values of V_{m2}, the latter occurring only when $\mathbf{p}_1, \cdots, \mathbf{p}_m$ lie in a semicircle, the above expected value implies the well known value of $m(\frac{1}{2})^{m-1}$ for the probability of this event.

The above calculations suggest that, for fixed n, EV_{mn} is asymptotically of the form $K_n m$. This section is devoted to finding upper and lower bounds on EV_{mn}, which establish this asymptotic growth rate and bound the growth of K_n for large n. (Unfortunately, the tempting conjecture that $K_n = n-1$ is false: K_n is approximately of order $(\beta n)^{n/2}$, as we have noted.)

First, we will establish an upper bound by showing that, for $2 \leqq n < m$,

$$EV_{mn} \leqq A_n m(1 - B_{mn}), \tag{4.1}$$

where

$$A_n = c(n^2 - 2n)\left(\frac{2}{c(n-2)}\right)^{n-1} \frac{(n-1)^{n-3}}{n},$$

$$B_{mn} = \left(\frac{1}{2}\right)^{m-1} \sum_{j=0}^{n-2} \binom{m-1}{j} \left(\frac{c(n-2)}{n-1}\right)^j.$$

Before proceeding, we remark that B_{mn} is negligible:

$$B_{mn} < \left(\frac{1}{2}\right)^{m-1} \left(1 + \frac{c(n-2)}{n-1}\right)^{m-1},$$

and since $(c(n-2)/(n-1)) < \frac{3}{4}$ when $n \geqq 3$, $B_{mn} \leqq (\frac{7}{8})^{m-1}$.

We observe also that Wallis's formula [1, p. 238], viz.

$$\frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(\alpha + 1)} \sim \frac{1}{\sqrt{\alpha}}\left(1 - \frac{1}{8\alpha} + O\left(\frac{1}{\alpha^2}\right)\right) \quad \text{as } \alpha \to \infty,$$

implies

$$c(n^2 - 2n) \sim \sqrt{2/\pi}\, n$$

and

$$\frac{2}{c(n-2)} \sim \sqrt{2\pi/(n-2)}\left(1 - \frac{1}{4(n-2)}\right).$$

Thus

$$A_n \sim \frac{(2\pi)^{n/2}}{\pi e^{1/4}} n^{(n-5)/2},$$

and so if $m > n$, then asymptotically as $n \to \infty$,

$$EV_{mn} \leqq \frac{(2\pi)^{n/2}}{\pi e^{1/4}} n^{(n-5)/2} m. \tag{4.2}$$

The proof of (4.1) begins with (3.3), which says that

$$EV_{mn} = \binom{m}{n} c(n^2 - 2n) T_{n-2, m-n, n(n-2)} \tag{4.3}$$

where

$$T_{j,k,l} = \int_0^{\pi/2} (1 - \tfrac{1}{2}F_j(r))^k \sin^l r \, dr. \tag{4.4}$$

If $l \geqq j+1$, an integration by parts using

$$u = -\frac{2}{c(j)} \sin^{l-j} r, \qquad dv = (1-\tfrac{1}{2}F_j(r))^k (-\tfrac{1}{2}c(j) \sin^j r)\, dr$$

yields

$$T_{j,k,l} = \frac{-2}{c(j)} \frac{1}{k+1} \left(\frac{1}{2}\right)^{k+1} + \frac{2}{c(j)} \frac{l-j}{k+1} \int_0^{\pi/2} (1-\tfrac{1}{2}F_j(r))^{k+1} \sin^{l-j-1} r \cos r\, dr,$$

and hence

$$T_{j,kl} \leqq -\frac{2}{c(j)} \frac{1}{k+1} \left(\frac{1}{2}\right)^{k+1} + \frac{2}{c(j)} \frac{l-j}{k+1} T_{j,k+1,l-j-1} \quad \text{for } l \geqq j+1. \tag{4.5}$$

Iterating (4.5) gives, for $l \geqq \nu(j+1)$,

$$\begin{aligned} T_{j,k,l} \leqq & -\frac{2}{c(j)} \frac{1}{k+1} \left(\frac{1}{2}\right)^{k+1} \\ & - \sum_{i=2}^{\nu} \left(\frac{2}{c(j)}\right)^i \frac{(l-j)(l-2j-1)\cdots(l-(i-1)j-(i-2))}{(k+1)(k+2)\cdots(k+i)} \left(\frac{1}{2}\right)^{k+1} \\ & + \left(\frac{2}{c(j)}\right)^{\nu} \frac{(l-j)(l-2j-1)\cdots(l-\nu j-(\nu-1))}{(k+1)(k+2)\cdots(k+\nu)} T_{j,k+\nu,l-\nu(j+1)}. \end{aligned} \tag{4.6}$$

In particular, this is valid for $j = n-2$, $k = m-n$, $l = n(n-2)$, and $\nu = n-2$. In this case, for $i = 2, 3, \cdots, n-2$ we have

$$\begin{aligned} &(l-j)(l-2j-1)\cdots(l-(i-1)j-(i-2)) \\ &\quad = [(n-1)(n-2)][(n-1)(n-3)]\cdots[(n-1)(n-i)] \\ &\quad = (n-1)^{i-1} \frac{(n-2)!}{(n-i-1)!}, \end{aligned}$$

and also by a simple integration,

$$\begin{aligned} T_{j,k+\nu,l-\nu(j+1)} &= T_{n-2,m-2,n-2} \\ &= \int_0^{\pi/2} (1-\tfrac{1}{2}F_{n-2}(r))^{m-2} \sin^{n-2} r\, dr \\ &= \frac{2}{c(n-2)} \frac{1}{m-1} (1-(\tfrac{1}{2})^{m-1}). \end{aligned}$$

Putting the two expressions above into (4.6) yields

$$\begin{aligned} T_{n-2,m-n,n(n-2)} \leqq & \left(\frac{2}{c(n-2)}\right)^{n-1} \frac{(n-2)!\,(m-n)!}{(m-1)!} (n-1)^{n-2} \\ & - \sum_{i=1}^{n-1} \left(\frac{2}{c(n-2)}\right)^i \frac{(n-2)!\,(m-n)!}{(n-i-1)!\,(m-n+i)!} (n-1)^{i-1} (\tfrac{1}{2})^{m-n+i}. \end{aligned}$$

Multiplying by $c(n^2-2n)\binom{m}{n}$ gives an upper bound on EV_{mn}; simplifying and letting the sum run over $j = n-1-i$ then gives (4.1).

Next we show that, for any $n \geqq 2$ and $\varepsilon \in (0, 1)$,

$$EV_{mn} \geqq \frac{d((n-1)^2)}{d(n-1)^{n-1}} \frac{2^{n-1}\varepsilon^n}{n!e^{n-1}}(n-1)^{n-1}m, \quad \text{asymptotically as } m \to \infty, \tag{4.7}$$

where

$$d(j) = \frac{\Gamma(\frac{1}{2}j + \frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2}j+1)}.$$

Before proceeding, we notice that Wallis's formula implies

$$d(j) \sim \sqrt{2/\pi j} \quad \text{as } j \to \infty.$$

This and Stirling's formula provide that

$$\frac{d((n-1)^2)}{d(n-1)^{n-1}} \frac{2^{n-1}\varepsilon^n}{n!e^{n-1}}(n-1)^{n-1} \sim (2\pi)^{(n-1)/2}\varepsilon^n n^{(n-6)/2} \quad \text{as } n \to \infty.$$

Therefore, for fixed large n, (4.7) implies that for any $\varepsilon \in (0, 1)$ we have, asymptotically as $m \to \infty$,

$$E_{mn}V \geqq \frac{(2\pi)^{(n-1)/2}}{\pi}\varepsilon^n n^{(n-6)/2}m. \tag{4.8}$$

To prove (4.7), we begin again with (3.3). For any $\alpha \in (0, \frac{1}{2}\pi)$,

$$\begin{aligned} EV_{mn} &\geqq \binom{m}{n} \int_0^\alpha (1 - \tfrac{1}{2}F_{n-2}(\alpha))^{m-n} f_{n(n-2)}(r)\, dr \\ &= \binom{m}{n}(1 - \tfrac{1}{2}F_{n-2}(\alpha))^{m-n} F_{n(n-2)}(\alpha). \end{aligned} \tag{4.9}$$

It can be checked that, for $0 \leqq \alpha < \frac{1}{2}\pi$,

$$F_k(\alpha) = \cos\alpha\,(d(k+1)\sin^{k+1}\alpha + d(k+3)\sin^{k+3}\alpha + \cdots)$$

by writing $F_k(\alpha)$ as incomplete beta function (through the substitution $s = \sin^2 r$) and then using a series expansion like that in [1, p. 944]. Moreover, $d(0) \geqq d(1) \geqq d(2) \geqq \cdots$. Therefore,

$$\begin{aligned} 1 - \tfrac{1}{2}F_{n-2}(\alpha) &\geqq 1 - \tfrac{1}{2}\cos\alpha\,(d(n-1)(\sin^{n-1}\alpha + \sin^{n+1}\alpha + \cdots)) \\ &= 1 - \frac{d(n-1)\sin^{n-1}\alpha}{2\cos\alpha} \end{aligned}$$

and

$$F_{n(n-2)}(\alpha) \geqq d((n-1)^2)\cos\alpha\,\sin^{(n-1)^2}\alpha.$$

So, writing τ for $\sin\alpha$, we have

$$EV_{mn} \geqq \binom{m}{n} d((n-1)^2)\tau^{(n-1)^2}\sqrt{1+\tau^2}\left(1 - \frac{d(n-1)}{2}\frac{\tau^{n-1}}{\sqrt{1-\tau^2}}\right)^{m-n} \quad \text{for any } \tau \in (0, 1).$$

Let ε be an arbitrary number in $(0, 1)$; then we have

$$EV_{mn} \geqq \binom{m}{n} d((n-1)^2)\tau^{(n-1)^2}\varepsilon\left(1 - \frac{d(n-1)}{2\varepsilon}\tau^{n-1}\right)^{m-n}$$

$$\text{if } 0 < \varepsilon < 1 \text{ and } 0 < \tau < \sqrt{1-\varepsilon^2}.$$

If we now write σ for $(d(n-1)/2\varepsilon)\,\tau^{n-1}$, then this inequality becomes

$$EV_{mn} \geqq \binom{m}{n} d((n-1)^2)\left(\frac{2\varepsilon\sigma}{d(n-1)}\right)^{n-1} \varepsilon(1-\sigma)^{m-n}$$

$$\text{if } 0<\varepsilon<1 \text{ and } 0<\sigma<\frac{d(n-1)}{2\varepsilon}(1-\varepsilon^2)^{(n-1)/2}.$$

Notice that the value of σ maximizing the right side of the above inequality is $(n-1)/(m-1)$. So if $(n-1)/(m-1)<(d(n-1)/2\varepsilon)(1-\varepsilon^2)^{(n-1)/2}$, then

$$EV_{mn} \geqq \left(\frac{m}{n}\right)\frac{d((n-1)^2}{d(n-1)^{n-1}}2^{n-1}\varepsilon^n\left(\frac{n-1}{m-1}\right)^{n-1}\left(1-\frac{n-1}{m-1}\right)^{m-n}.$$

Since $\binom{m}{n} \geqq m(m-n)^{n-1}/n!$,

$$EV_{mn} \geqq m\frac{d((n-1)^2)}{d(n-1)^{n-1}}\frac{2^{n-1}\varepsilon^n}{n!}\left(1-\frac{n-1}{m-1}\right)^{m-n}(n-1)^{n-1},$$

and (4.7) follows from this.

5. Comparison with known extreme values. The values of EV_{mn} can be compared to the maximum and minimum numbers of vertices for an n-dimensional polyhedron with m facets. For bounded polyhedra, these numbers are (Klee [5]):

$$\max V_{mn} = \binom{m-\left[\frac{n+1}{2}\right]}{m-n} + \binom{m-\left[\frac{n+2}{2}\right]}{m-n},$$

$$\min V_{mn} = (m-n)(n-1)+2.$$

For unbounded polyhedra, the minimum is significantly smaller, namely $m-n+1$. However, for our generation method, the probability that a random polyhedron is bounded is greater than $\frac{1}{2}$ when $m>2n$ and, for any fixed value of n, increases rapidly with m [6]. For the choices of m and n in the tables below, the larger minimum provides a better comparison.

Unfortunately, the integral formula (3.3) for EV_{mn} is difficult to evaluate numerically, even for moderate m and n. Therefore, the larger values of EV_{mn} in the accompanying tables may be somewhat inaccurate and should be taken to indicate the order of magnitude of EV_{mn} rather than the precise value. The numerical integration was done by means of Gaussian quadrature routine.

In Tables 1 and 2, values of EV_{mn} are given with n fixed at 4 and 7, respectively. The essential asymptotic linearity in m, predicted by the estimates (4.2) and (4.8), is evident. Note that the maximum number of vertices has the asymptotic growth $\max V_{mn} \sim m^{n/2}$ for fixed n.

In Tables 3 and 4, values of EV_m when $m=2n$ and $m=5n$ are given as n varies between 2 and 15. These figures illustrate the rapid increase of EV_{mn} with n, although the increase is slower than that of $\max V_{mn}$. Asymptotic estimates of EV_{mn} as $n\to\infty$ for various fixed rates of growth of m as a function of n would provide greater insight into the structure of randomly generated polyhedra.

TABLE 1
Values for the case $n = 4$

m	min V_{m4}	EV_{m4}	max V_{m4}
10	20	25	35
20	50	79	170
30	80	137	405
40	110	197	740
50	140	259	1175
100	290	574	4850
110	320	638	5885
120	350	703	7020
130	380	767	8255
140	410	832	9590
150	440	897	11025
500	1490	3204	124250
510	1520	3271	129285
520	1550	3337	134420
530	1580	3404	139655
540	1610	3470	144990
550	1640	3537	150425
1000	2990	6540	498500
1010	3020	6607	508535
1020	3050	6674	518670
1030	3080	6741	528905
1040	3110	6808	539240
1050	3140	6875	549675

TABLE 2
Values for the case $n = 7$

m	min V_{m7}	EV_{m7}	max V_{m7}
20	80	658	1120
30	140	2040	5200
40	200	4033	14280
50	260	6518	30360
100	560	23982	285760
110	620	28205	385840
120	680	32614	506920
130	740	37193	651000
140	800	41926	820080
150	860	46802	1016160

TABLE 3
Values for the case $m = 2n$

n	m	min V_{mn}	EV_{mn}	max V_{mn}
2	4	4	3.5	4
3	6	8	7.1	8
4	8	14	15.5	20
5	10	22	35	42
6	12	32	79	112
7	14	44	180	240
8	16	58	416	660
9	18	74	966	1430
10	20	92	2251	4004
11	22	112	5267	8736
12	24	134	12359	24752
13	26	158	29080	54264
14	28	184	68581	155040
15	30	212	162073	341088

TABLE 4
Values for the case $m = 5n$ (rounded to 3 significant digits)

n	m	min V_{mn}	EV_{mn}	max V_{mn}
2	10	10	10	10
3	15	26	26	26
4	20	50	79	170
5	25	82	258	462
6	30	122	867	3250
7	35	170	2970	8990
8	40	226	10300	65500
9	45	290	36000	183000
10	50	362	127000	1360000
11	55	446	450000	3810000
12	60	530	1600000	34400000
13	65	626	5730000	80900000
14	70	730	20600000	615000000
15	75	842	74000000	1740000000

REFERENCES

[1] M. ABRAMOWITZ AND I. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1965.
[2] J. DUNHAM, D. KELLY AND J. TOLLE, *Some experimental results concerning the expected number of pivots for solving randomly generated linear programs*, Tech. Rep. 77-16, ORSA Curriculum, Univ. of North Carolina at Chapel Hill, 1977.
[3] W. FELLER, *An Introduction to Probability Theory and Its Applications*, vol. 2, 2nd ed., John Wiley, New York, 1971.
[4] D. KELLY AND J. TOLLE, *Expected simplex algorithm behaviour for random linear programs*, Oper. Res. Verfahren 31 (1979), pp. 361–367.
[5] V. KLEE, *Polytope pairs and their relationship to linear programming*. Acta Math. 133 (1974), pp. 1–25.
[6] T. MATTHEISS AND B. SCHMIDT, *The probability that a random polytope is bounded*, Math. Oper. Res. 2 (1977), pp. 292–296.
[7] R. MILES, *Random points, sets, and tessellations on the surface of a sphere*, Sankhyā Ser. A, 33 (1971), pp. 145–174.

[8] A. PRÉKOPA, *On the number of vertices of random convex polyhedra*, Period. Math. Hungar. 2 (1972), pp. 259–282.

[9] A. RENYI AND R. SULANKE, *Zufällige konvexe Polygone in einem Ringgebiet*, Z. Wahrsch. Verw. Gebiete 9 (1968), pp. 146–157.

[10] W. SCHMIDT, *Some results in probabilistic geometry*. Z. Wahrsch. Verw. Gebiete 9 (1968), pp. 158–162.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 4, December 1981

0196-5212/81/0204-0011 $01.00/0

WEIGHT ENUMERATORS OF SELF-ORTHOGONAL CODES OVER GF(3)*

C. L. MALLOWS† AND N. J. A. SLOANE†

Abstract. The Hamming and complete weight enumerators of maximally self-orthogonal codes over $GF(3)$ of lengths $12m-1$, $12m$ and $12m+1$ are characterized. The results for length $12m+1$ are believed to be new, while those for length $12m-1$ and $12m$ have been considerably simplified.

1. Introduction. Professor Marshall Hall, Jr., recently pointed out to us (in connection with his work on the hypothetical projective plane of order twelve) that there is an omission in [6]: the maximally self-orthogonal ternary codes of length $12m+1$ are not mentioned. An important example of this class is the dual of the code generated by the incidence matrix of the projective plane of order 3, denoted by p_{13} in [1] and [10]. In fact it is incorrect to say (as we do in [6, p.655]) that if C is a maximally self-orthogonal $[n, \frac{1}{2}(n-1)]$ code over $GF(3)$ with $1 \in C^{\perp}$ then the extended code $(C^{\perp})^{+}$ is self-dual and n is congruent to -1 modulo 12. We shall see that the correct conclusion is that either $n \equiv -1 \pmod{12}$ and $(C^{\perp})^{+}$ is self-dual, or $n \equiv +1 \pmod{12}$ and $(C^{\perp})^{+}$ is not self-dual. The present paper gives the weight enumerators for the missing case. While determining these we were able to considerably simplify the weight enumerators in the case $n \equiv -1 \pmod{12}$ and also when C itself is self-dual and $n \equiv 0 \pmod{12}$. One might say that this is an error-correcting paper.

2. Weight enumerators. Let C be a code of length n and dimension k over $GF(3)$. The complete weight enumerator (cwe) of C is

$$E_C(x, y, z) = \sum_{\mathbf{u} \in C} x^{n_0(\mathbf{u})} y^{n_1(\mathbf{u})} z^{n_2(\mathbf{u})},$$

where $n_i(\mathbf{u})$ is the number of components of $\mathbf{u}$ that are congruent to i modulo 3. The ordinary or Hamming weight enumerator of C is

$$W_C(x, y) = E_C(x, y, y).$$

From the MacWilliams theorem the complete weight enumerator of the dual code $C^{\perp}$ is given by

$$E_{C^{\perp}}(x, y, z) = \frac{1}{3^k} E_C(x+y+z, x+\omega y+\omega^2 z, x+\omega^2 y+\omega z), \tag{1}$$

where $\omega = e^{2\pi i/3}$ (see [4], [6]).

A code C is called self-orthogonal if $C \subseteq C^{\perp}$ and self-dual if $C = C^{\perp}$. The maximum dimension of a self-orthogonal code of length n is $\frac{1}{2}n$ if $n \equiv 0 \pmod 4$, $\frac{1}{2}(n-2)$ if $n \equiv 2 \pmod 4$, and $\frac{1}{2}(n-1)$ if n is odd (see [8], [9]). We wish to characterize the Hamming and complete weight enumerators of self-orthogonal codes of maximal dimension. However our method will only work when we can express the cwe of $C^{\perp}$ in terms of the cwe of C. There are two general cases when we can do this:

(i) when C is self-dual, so that

$$E_{C^{\perp}}(x, y, z) = E_C(x, y, z), \tag{2}$$

* Received by the editors October 27, 1980.
† Bell Laboratories, Murray Hill, New Jersey 07974.

(ii) when C is maximally self-orthogonal of odd length and the all-ones vector $\mathbf{1}$ is in $C^\perp$ but not in C.

In case (ii) we have

$$\dim C^\perp = \dim C + 1 = \tfrac{1}{2}(n+1),$$

which implies

$$C^\perp = C \cup (\mathbf{1}+C) \cup (\mathbf{2}+C) \tag{3}$$

and

$$W_{C^\perp}(x, y, z) = W_C(x, y, z) + W_C(y, z, x) + W_C(z, x, y) \tag{4}$$

(since the cwe of $\mathbf{1}+C$ is $W_C(y, z, x)$, etc.).

In case (i) the length must be a multiple of 4, and if we make the additional assumption that the all-ones vector is in C then $n \equiv 0 \pmod{12}$. Without this assumption the results are far more complicated (see [6]). In case (ii) it is a consequence of Theorem 3 below that $n \equiv \pm 1 \pmod{12}$. If $n \equiv -1$ we can add an overall parity check to $C^\perp$ so as to make $C^\perp$ self-dual, but if $n \equiv +1 \pmod{12}$ this is impossible.

In order to describe the weight enumerators of these codes we introduce the following polynomials. We apologize for the length of this, but it is essential for our method that we work with homogeneous polynomials in six variables. As far as possible we use the same notation as [6].

First the polynomials in x, y, z:

$$\begin{aligned}
a &= x^3+y^3+z^3,\\
f &= x^2y+y^2z+z^2x,\\
g &= xy^2+yz^2+zx^2,\\
p &= 3xyz,\\
\psi_4 &= x(x^3+8y^3),\\
\phi_4 &= y(x^3-y^3),\\
\xi_4 &= x(y^3-z^3),\\
b &= x^3y^3+y^3z^3+z^3x^3,\\
\beta_6 &= a^2-12b\\
&= x^6+y^6+z^6-10(x^3y^3+y^3z^3+z^3x^3),\\
\upsilon_7 &= x(2x^6-7y^6-7z^6+7x^3y^3+7x^3z^3-56y^3z^3),\\
\pi_9 &= (x^3-y^3)(y^3-z^3)(z^3-x^3),\\
\alpha_{12} &= a(a^3+8p^3)\\
&= \sum^{(3)} x^{12}+4\sum^{(6)} x^9y^3+6\sum^{(3)} x^6y^6+228\sum^{(3)} x^6y^3z^3,\\
\tau_{13} &= xy^6(x^3-y^3)(2x^3+y^3).
\end{aligned}$$

The second set are polynomials in u, v, w, x, y, z:

$$\Lambda_2 = ux + vy + wz,$$

$$\Xi_5 = ux(y^3 - z^3) + vy(z^3 - x^3) + wz(x^3 - y^3),$$

$$\begin{aligned}\Upsilon_8 = {} & ux(2x^6 - 7y^6 - 7z^6 + 7x^3y^3 + 7x^3z^3 - 56y^3z^3) \\ & + vy(2y^6 - 7z^6 - 7x^6 + 7y^3z^3 + 7x^3y^3 - 56x^3z^3) \\ & + wz(2z^6 - 7x^6 - 7y^6 + 7x^3z^3 + 7y^3z^3 - 56x^3y^3).\end{aligned}$$

Note that

$$\pi_9 = g^3 - f^3,$$

$$243\tau_{13} = x\psi_4^3 - 37x\phi_4^3 - \tfrac{1}{2}\beta_6 \upsilon_7\big|_{y=z},$$

$$\xi_4 = \Xi_5(u = 1, v = w = 0),$$

$$\upsilon_7 = \Upsilon_8(u = 1, v = w = 0).$$

We can now state our results.

THEOREM 1 (Complete weight enumerator). *If* $C = C^\perp$ *and* $\mathbf{1} \in C$, *then* $n \equiv 0 \pmod{12}$ *and*

$$E_C(x, y, z) \in R \oplus \beta_6 \pi_9^2 R, \tag{5}$$

where

$$R = \mathbf{C}[\beta_6^2, \alpha_{12}, \pi_9^4].$$

In other words the cwe *of* C *can be written uniquely as a polynomial in* β_6^2, α_{12} *and* π_9^4, *plus* $\beta_6\pi_9^2$ *times another such polynomial.*

COROLLARY 2 (Hamming weight enumerator—Gleason [3]). *With the same hypotheses as Theorem* 1,

$$W_C(x, y) \in \mathbf{C}[\psi_4^3, \phi_4^3]. \tag{6}$$

THEOREM 3 (Complete weight enumerator). *If* $C \subset C^\perp$, $\mathbf{1} \in C^\perp \backslash C$, *and* $\dim C^\perp = \dim C + 1$ *then* $n \equiv \pm 1 \pmod{12}$.

(a) *If* $n = 12m + 1$ *then*

$$E_C(x, y, z) \in xR \oplus \beta_6\upsilon_7 R \oplus \xi_4\pi_9 R \oplus \upsilon_7\pi_9^2 R \oplus x\beta_6\pi_9^2 R \oplus \xi_4\beta_6\pi_9^3 R, \tag{7}$$

where R *is defined in Theorem* 1.

(b) *If* $n = 12m - 1$ *then*

$$E_C(x, y, z) \in \bar{\alpha}_{12} R \oplus \bar{\beta}_6\beta_6 R \oplus \bar{\beta}_6\pi_9^2 R \oplus \beta_6\bar{\pi}_9\pi_9 R \oplus \bar{\pi}_9\pi_9^3 \oplus \beta_6\pi_9^2\bar{\alpha}_{12} R, \tag{8}$$

where the bar denotes partial differentiation with respect to x.

COROLLARY 4 (Hamming weight enumerator). *With the same hypotheses as Theorem* 3,

(a) *if* $n = 12m + 1$ *then*

$$W_C(x, y) \in x\mathbf{C}[\psi_4^3, \phi_4^3] \oplus \tau_{13}\mathbf{C}[\psi_4^3, \phi_4^3], \tag{9}$$

and

(b) *if* $n = 12m - 1$ *then*

$$W_C(x, y) \in \bar{\psi}_4\psi_4^2\mathbf{C}[\psi_4^3, \phi_4^3] \oplus \bar{\phi}_4\phi_4^2\mathbf{C}[\psi_4^3, \phi_4^3]. \tag{10}$$

Extremal weight enumerators. Let us consider the extremal weight enumerators (as defined in [5] and [7]) corresponding to Theorem 3(a) and Corollary 4(a). The first nontrivial length is 13, and for simplicity we begin with the Hamming weight enumerator. By Corollary 4(a) the Hamming weight enumerator of any [13, 6] self-orthogonal code (with **1** in the dual code) has the form

$$W(x, y) = c_1 x\psi_4^3 + c_2 x\phi_4^3 + c_3\tau_{13} \tag{11}$$

for appropriate constants c_1, c_2, c_3. Suppose these constants are chosen so that the minimum weight of the corresponding code (if there is one) is as large as possible. We obtain

$$\begin{aligned} W(x, y) &= x\psi_4^3 - 24x\phi_4^3 - 132\tau_{13} \\ &= x^{13} + 572x^4y^9 + 156xy^{12} \\ &= W^* \qquad \text{(say)}. \end{aligned}$$

Since W^* has nonnegative integral coefficients, it could indeed be the weight enumerator of some [13, 6] code C^* with minimum weight 9. On the other hand from Theorem 3(a) the complete weight enumerator of C^* has the form

$$E_{C^*}(x, y, z) = b_1 x\beta_6^2 + b_2 x\alpha_{12} + b_3\beta_6\upsilon_7 + b_4\xi_4\pi_9 \tag{12}$$

for appropriate constants b_i. The condition that C^* has minimum weight 9 determines the b_i uniquely and we find

$$\begin{aligned} E_{C^*}(x, y, z) = x^{13} &+ 286x^4(y^6z^3 + y^3z^6) \\ &- \tfrac{13}{9}x(y^{12} + z^{12}) + \tfrac{286}{9}x(y^9z^3 + y^3z^9) + \tfrac{286}{3}xy^6z^6. \end{aligned}$$

Since this does not have nonnegative integral coefficients, C^* does not exist.

In this case we already knew from the enumeration in [1] that the highest minimum weight attainable is 6, and furthermore that there is a unique code with minimum weight 6, namely the projective plane code p_{13} mentioned in §1. For this code

$$\begin{aligned} W_{p_{13}}(x, y) &= x^{13} + 156x^7y^6 + 494x^4y^9 + 78xy^{12} \\ &= x\psi_4^3 - 24x\phi_4^3 - 54\tau_{13}, \end{aligned} \tag{13}$$

and

$$\begin{aligned} E_{p_{13}}(x, y, z) &= x^{13} + 156x^7y^3z^3 + 13x^4(y^9 + 18y^6z^3 + 18y^3z^6 + z^9) + 78xy^6z^6 \\ &= \tfrac{5}{72}x\beta_6^2 + \tfrac{17}{24}x\alpha_{12} + \tfrac{1}{9}\beta_6\upsilon_7 + 3\xi_4\pi_9 \end{aligned} \tag{14}$$

(which are of the forms (11) and (12)). Although the extremal weight enumerators did not tell us anything new in this example, it is nevertheless interesting to find a situation where the cwe leads to a contradiction not apparent from the Hamming weight enumerator.

For codes of greater length the extremal cwe will probably always contain a negative coefficient (compare [5] and [7]).

Relationship with [6]. The basis for cwe's given in Theorem 1 is simpler than that given in [6, Thm. 1]. The old basis is expressed in terms of the new one by

$$\gamma_{18} = -\tfrac{1}{2}\beta_6^3 + \tfrac{3}{2}\beta_6\alpha_{12} + 216\pi_9^2,$$

and

$$256\delta_{36} = -\beta_6^6 + 6\beta_6^4\alpha_{12} - 9\beta_6^2\alpha_{12}^2 + 4\alpha_{12}^3 + 864\beta_6^3\pi_9^2 - 2592\beta_6\pi_9^2\alpha_{12} - 186624\pi_9^4.$$

The syzygy $\gamma_{18}^2 = \alpha_{12}^3 - 64\delta_{36}$ has been replaced by the trivial identity

$$(\beta_6\pi_9^2)^2 = \beta_6^2 \cdot \pi_9^4.$$

3. The proofs.

Proof of Theorem 1. Since this is parallel to the proofs given in [11] and [12] our treatment will be brief. Suppose C is a self-dual code of length $n = 12m$ containing $\mathbf{1}$. Let G be the subgroup of $GL(3, \mathbf{C})$ of order 2952 generated by the matrices

$$M = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 & 1 & 1 \\ 1 & \omega & \omega^2 \\ 1 & \omega^2 & \omega \end{bmatrix}, \qquad J = \begin{bmatrix} 1 & & \\ & \omega & \\ & & 1 \end{bmatrix},$$

and all 3×3 permutation matrices. The matrices in G are listed below, in the proof of Lemma 6. In [11] and [12] it is shown that the cwe $E_C(x, y, z)$ is invariant under G. If a_n denotes the number of linearly independent homogeneous invariants for G of degree n, it is also shown in [11] and [12] that the Molien series $\sum_{n=0}^{\infty} a_n\lambda^n$ is equal to

$$\frac{1+\lambda^{24}}{(1-\lambda^{12})^2(1-\lambda^{36})}. \tag{15}$$

To find a basis for the invariants we proceed as follows. Let $s = e^{2\pi i/12}$. Under the action of M we have

$$\beta_6 \xrightarrow{M} -\beta_6, \qquad \alpha_{12} \xrightarrow{M} \alpha_{12}, \qquad x^3 - y^3 \xrightarrow{M} s^{11}(f - \omega^2 g) \xrightarrow{M} -(z^3 - x^3),$$

$$y^3 - z^3 \xrightarrow{M} s^3(f - g) \xrightarrow{M} -(y^3 - z^3), \qquad z^3 - x^3 \xrightarrow{M} s^7(f - \omega g) \xrightarrow{M} -(x^3 - y^3).$$

Therefore

$$\pi_9 = (x^3 - y^3)(y^3 - z^3)(z^3 - x^3) = g^3 - f^3 \xrightarrow{M} i\pi_9.$$

All of β_6, π_9 and α_{12} are invariant under J, and the permutations fix β_6 and α_{12} and send π_9 to $\pm\pi_9$. Thus β_6^2, α_{12}, $\beta_6\pi_9^2$ and π_9^4 are indeed invariant under G. It only remains to show that β_6^2, α_{12} and π_9^4 (or equivalently β_6, α_{12} and π_9) are algebraically independent. This is verified by computing the Jacobian of β_6, α_{12} and π_9, which is

$$-2592x^2y^2z^{20} + \cdots.$$

Since this does not vanish, the polynomials are indeed algebraically independent [2, Thm. 2.3]. Therefore the ring of invariants of G has the form shown on the right-hand side of (5). This completes the proof of Theorem 1.

Proof of Corollary 2. We set $y = z$ in Theorem 1, making π_9 vanish, and then replace β_6^2 and α_{12} by the equivalent but simpler pair ψ_4^3 and ϕ_4^3.

Proof of Theorem 3. Suppose C is an $[n, \frac{1}{2}(n-1)]$ self-orthogonal code with $\mathbf{1} \in C^{\perp}\backslash C$. This implies that the cwe of C is a polynomial in x, y^3 and z^3, and also satisfies

(4). From the MacWilliams identity (1) we have

$$\begin{aligned} M \circ E_C(x, y, z) &= 3^{-n/2} E_C(x+y+z, x+\omega y+\omega^2 y, x+\omega^2 y+\omega z) \\ &= 3^{-1/2} E_{C^\perp}(x, y, z) \\ &= 3^{-1/2}\{E_C(x, y, z)+E_C(y, z, x)+E_C(z, x, y)\}, \end{aligned} \tag{16}$$

using (4). Also

$$\begin{aligned} M \circ E_{1+C}(x, y, z) &= M \circ E_C(y, z, x) \\ &= 3^{-n/2} E_C(x+\omega y+\omega^2 z, x+\omega^2 y+\omega z, x+y+z) \\ &= 3^{-1/2} E_{C^\perp}(x, \omega y, \omega^2 z) \qquad \text{(from (16))} \\ &= 3^{-1/2}\{E_C(x, \omega y, \omega^2 z)+E_C(\omega y, \omega^2 z, x)+E_C(\omega^2 z, x, \omega y)\} \\ &= 3^{-1/2}\{E_C(x, y, z)+\omega^n E_C(y, z, x)+\omega^{2n} E_C(z, x, y)\}, \end{aligned} \tag{17}$$

where in the last step we used the fact that $n_0(\mathbf{u}) \equiv n \pmod 3$ and $n_1(\mathbf{u}) \equiv n_2(\mathbf{u}) \equiv 0 \pmod 3$ for all $\mathbf{u} \in C$. Similarly,

$$M \circ E_C(z, x, y) = 3^{-1/2}\{E_C(x, y, z)+\omega^{2n} E_C(y, z, x)+\omega^n E_C(z, x, y)\}. \tag{18}$$

Since n is odd we have to consider the possibilities $n \equiv \pm 1, \pm 3$ and $\pm 5 \pmod{12}$.

Case 1. $n \equiv \pm 3 \pmod{12}$. The last expressions in (16) and (17) are now identical, implying $E_C(x, y, z) = E_C(y, z, x)$. Since C always contains $\mathbf{0}$, this implies that $\mathbf{1} \in C$, a contradiction. Thus $n \not\equiv \pm 3 \pmod{12}$.

Case 2. $n \equiv +1$ or $-5 \pmod{12}$. Now (17) becomes

$$M \circ E_C(y, z, x) = 3^{-1/2}\{E_C(x, y, z)+\omega E_C(y, z, x)+\omega^2 E_C(z, x, y)\}.$$

We introduce new indeterminates u, v, w and define

$$F(u, v, w, x, y, z) = uE_C(x, y, z)+vE_C(y, z, x)+wE_C(z, x, y).$$

Let G^* denote the group of 6×6 matrices

$$\left\{A^* = \begin{pmatrix} \bar{A} & 0 \\ 0 & A \end{pmatrix}; A \in G\right\}$$

of order 2592, where $\bar{A}$ acts on the variables u, v, w and A acts on x, y, z. From (16), (17) and (18) it follows that F is invariant under M^* and thus under all of G^*. (Compare the proof of [6, Thm. 9].) We indicate this by writing

$$F(\bar{A}\mathbf{u}, A\mathbf{x}) = F(\mathbf{u}, \mathbf{x}) \quad \text{all } A \in G. \tag{19}$$

At this point we need the following analogue of [6, Thm. 8]. (The proof is essentially the same and is omitted.)

THEOREM 5. *Let G be any finite subgroup of $GL(m, \mathbf{C})$, and let Φ_d denote the set of all polynomials $F(\mathbf{u}, \mathbf{x}) = F(u_1, \cdots, u_m, x_1, \cdots, x_m)$ which are*

(i) *homogeneous of total degree d,*

(ii) *linear in the u_i, and*

(iii) *satisfy* (19).

Let a_d denote the number of linearly independent polynomials in Φ_d. Then a generating function for the numbers a_d is

$$\sum_{d=0}^{\infty} a_d \lambda^d = \frac{\lambda}{|G|} \sum_{A \in G} \frac{\operatorname{tr}(\bar{A})}{\det(I - \lambda A)}. \tag{20}$$

The next step is to compute the sum on the right-hand side of (20) for our group.

LEMMA 6. *If G is the subgroup of $GL(3, \mathbf{C})$ of order 2592 defined at the beginning of this section,*

$$\frac{\lambda}{|G|}\sum_{A\in G}\frac{\operatorname{tr}(\bar{A})}{\det(I-\lambda A)}=\lambda\cdot\frac{\lambda+\lambda^{13}}{(1-\lambda^{12})^3} \tag{21}$$
$$=\lambda\cdot\frac{\lambda+2\lambda^{13}+2\lambda^{25}+\lambda^{37}}{(1-\lambda^{12})^2(1-\lambda^{36})}.$$

Proof. We know from [11] and [12] that there are (I) 1944 elements of G of the form

$$\begin{bmatrix} & 1 & \\ s^\nu & & \omega^a \\ & & & \omega^b \end{bmatrix} M^e \begin{bmatrix} 1 & & \\ & \omega^c & \\ & & \omega^d \end{bmatrix}$$

and (II) 648 elements of the form

$$\begin{bmatrix} & 1 & \\ s^\nu & & \omega^a \\ & & & \omega^b \end{bmatrix} P,$$

where $0\leqq\nu\leqq 11$, $0\leqq a, b, c, d\leqq 2$, $e=1$ or 3 and P is any 3×3 permutation matrix. Instead of (20) we shall work out

$$\frac{1}{\lambda}\sum_{A\in G}\frac{\operatorname{tr}(\bar{A})}{\det(I-\lambda A)}=\sum_{A\in G}\frac{\operatorname{tr}(\lambda A)^{-1}}{\det(I-\lambda A)}$$

(since G is a unitary group). Our strategy is to keep $\nu=0$ as long as possible, finally summing over ν by replacing λ by λs^ν and adding. For type (I) we can ignore c and d (and just multiply the final sum by 9) since c can be combined with a and d with b in both $\operatorname{tr}(\lambda A)^{-1}$ and $\det(I-\lambda A)$. The sum on a, b and e is equal to

$$\frac{2}{1-\lambda^4}+\frac{2\sqrt{3}/\lambda+6^*+\sqrt{3}\lambda+3\lambda^2}{1+\lambda^6}+\frac{2\sqrt{3}/\lambda+4^*+2\sqrt{3}\lambda^3+2\lambda^4}{(1+\lambda^2+\lambda^4)(1-\lambda^2+\lambda^4)}$$
$$+\frac{2\sqrt{3}/\lambda+6^*+\sqrt{3}\lambda-3\lambda^2-\sqrt{3}\lambda^3-3\lambda^{4^*}-3\sqrt{3}\lambda^5-3\lambda^{6^*}}{(1-\lambda^2+\lambda^4)(1+\lambda^6)}.$$

We replace λ by λs^ν and sum over ν; to do this we put everything over powers of $1-\lambda^{12}$ and ignore all terms that are not powers of λ^{12} (those not marked with an asterisk). This gives

$$\frac{24}{1-\lambda^{12}}+\frac{72}{1-\lambda^{12}}+\frac{48}{1-\lambda^{12}}+\frac{12(6+6\lambda^{12}+6\lambda^{12}+6\lambda^{12})}{(1-\lambda^{12})^2}$$

and multipling by 9 to account for the summation over c and d we find that the contribution from the type (I) terms is

$$\frac{9\cdot 144}{1-\lambda^{12}}+\frac{108(6+8\lambda^{12})}{(1-\lambda^{12})^2}. \tag{22}$$

Next we consider the type (II) terms. When $P = I$ the sum on a and b gives

$$\frac{9}{\lambda}\frac{1+2\lambda}{(1-\lambda)(1-\lambda^3)^2}. \tag{23}$$

For

$$P=\begin{bmatrix}1&0&0\\0&0&1\\0&1&0\end{bmatrix}$$

the contribution is

$$\frac{9}{\lambda}\frac{1}{(1-\lambda)(1-\lambda^6)}, \tag{24}$$

and for

$$P=\begin{bmatrix}0&1&0\\1&0&0\\0&0&1\end{bmatrix} \text{ or } \begin{bmatrix}0&0&1\\0&1&0\\1&0&0\end{bmatrix}$$

we get

$$\frac{9}{(1-\lambda^3)(1-\lambda^6)} \quad \text{(twice)}. \tag{25}$$

Finally

$$P=\begin{bmatrix}0&1&0\\0&0&1\\1&0&0\end{bmatrix} \text{ and } \begin{bmatrix}0&0&1\\1&0&0\\0&1&0\end{bmatrix}$$

have trace 0. Adding (23)–(25) we obtain

$$\frac{9}{\lambda}\frac{2+4\lambda-2\lambda^2+2\lambda^4}{(1-\lambda)(1-\lambda^3)(1-\lambda^6)},$$

and we sum over ν as before, by multiplying top and bottom by $(1+\lambda+\lambda^2)(1+\lambda^3)^2(1+\lambda^6)^3$, to get

$$\frac{9\cdot 12}{(1-\lambda^{12})^3}(6+36\lambda^{12}+6\lambda^{24}). \tag{26}$$

The grand total is the sum of (22) and (26):

$$2596\frac{1+\lambda^{12}}{(1-\lambda^{12})^3}.$$

We multiply by $\lambda^2/2596$ to obtain (21). The final step is to multiply the top and bottom by $1+\lambda^{12}+\lambda^{24}$ to make the denominator agree with that of (15). This completes the proof of Lemma 6.

Since only exponents of the form $12m+1$ appear in (21), we see that n cannot be of the form $12m-5$. It remains to find a basis for the sets Φ_d defined in Theorem 5. We compute

$$\Xi_5 \xrightarrow{M} -\Xi_5, \qquad \Upsilon_8 \xrightarrow{M} -\Upsilon_8$$

and deduce the following theorem from (21).

THEOREM 7. *The solutions of* (19) *that are linear in u, v, w belong to*

$$\Lambda_2 R \oplus \beta_6 \Upsilon_8 R \oplus \Xi_5 \pi_9 R \oplus \Upsilon_8 \pi_9^2 R \oplus \Lambda_2 \beta_6 \pi_9^2 R \oplus \Xi_5 \beta_6 \pi_9^3 R.$$

Finally, setting $u = 1$ and $v = w = 0$ in Theorem 7, we obtain (7) and thus prove Theorem 3(a).

Case 3. $n \equiv -1$ or $+5 \pmod{12}$. Now $F(u, v, w, x, y, z)$ is invariant under

$$G^{**} = \left\{ \begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix}; A \in G \right\}.$$

The proof in this case is essentially given in [6, p. 657].

This completes the proof of theorem 3.

Proof of Corollary 4. We set $y = z$ in Theorem 3, making ξ_4 and π_9 vanish, and replace $\beta_6 \upsilon_7$ by τ_{13}.

Remark. The Taylor series expansion of (21) is

$$\lambda \sum_{m=0}^{\infty} (m+1)^2 \lambda^{12m+1}.$$

Therefore the number of linearly independent homogeneous polynomials of degree $12m+1$ in the right-hand side of (7) is $(m+1)^2$. Similarly the number of degree $12m-1$ in (8) is $m(m+1)$.

Acknowledgments. We are grateful to Professor Marshall Hall Jr. for pointing out the omission in [6] which led to this work. During this investigation we have made use of two computer programs for symbolic manipulation: the MACSYMA system at the Massachusetts Institute of Technology Laboratory for Computer Science, and the ALTRAN system at the Bell Laboratories Murray Hill Computation Center.

REFERENCES

[1] J. H. CONWAY, V. PLESS AND N. J. A. SLOANE, *Self-dual codes over GF(3) and GF(4) of length not exceeding* 16, IEEE Trans. Information Theory, IT-25 (1979), pp. 312–322.

[2] L. FLATTO, *Invariants of finite reflection groups*, L'Enseignement mathématique, 24 (1978), pp. 237–292.

[3] A. M. GLEASON, *Weight polynomials of self-dual codes and the MacWilliams identities*, in Actes, Congrès International des Mathématiciens 1970, Vol. 3, Gauthiers–Villars, Paris, 1971, pp. 211–215.

[4] F. J. MACWILLIAMS AND N. J. A. SLOANE, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam and Elsevier/North-Holland, New York, 1977.

[5] C. L. MALLOWS, A. M. ODLYZKO AND N. J. A. SLOANE, *Upper bounds for modular forms, lattices and codes*, J. Algebra, 36 (1975), pp. 68–76.

[6] C. L. MALLOWS, V. PLESS AND N. J. A. SLOANE, *Self-dual codes over GF(3)*, SIAM J. Applied Math., 31 (1976), pp. 649–666.

[7] C. L. MALLOWS AND N. J. A. SLOANE, *An upper bound for self-dual codes*, Inform. and Control, 22 (1973), pp. 188–200.

[8] V. PLESS, *The number of isotropic subspaces in a finite geometry*, Rend. Cl. Scienze Fisiche, Matematiche e Naturali, Acc. Naz. Lincei, 39 (1969), pp. 418–421.

[9] ———, *On the uniqueness of the Golay codes*, J. Combinatorial Theory, 5 (1968), pp. 215–228.

[10] V. PLESS, N. J. A. SLOANE AND H. N. WARD, *Ternary codes of minimum weight* 6 *and the classification of the self-dual codes of length* 20, IEEE Trans. Information Theory, IT-26 (1980), pp. 305–316.

[11] N. J. A. SLOANE, *Weight enumerators of codes*, in Combinatorics, M. Hall Jr. and J. H. van Lint, eds., Reidel, Dordrecht and Mathematical Centre, Amsterdam, 1974, pp. 115–142.

[12] N. J. A. SLOANE, *Error-correcting codes and invariant theory: new applications of a nineteenth-century technique*, Amer. Math. Monthly, 84 (1977), pp. 82–107.

SIAM J. ALG. DISC. METH.
Vol. 2, No. 4, December 1981

0196-5212/81/0204-0012 $01.00/0

STRONG CONNECTIVITY IN DIRECTIONAL NEAREST-NEIGHBOR GRAPHS*

B. E. FLINCHBAUGH† AND L. K. JONES‡

Abstract. A Directional Nearest-Neighbor graph is defined on a finite set of points in the plane by drawing an arc from each point X to its nearest neighbor in each of r divisions of the plane relative to X. We prove that Directional Nearest-Neighbor graphs having $r = 4$ are strongly connected.

1. Introduction. The problem of representing geographic information for subsequent analysis by computer has motivated an interesting class of graphs which we call directional nearest-neighbor graphs. Potential applications of the representation lie in problems in forestry, dynamic monitoring of water table levels, and other environmental studies requiring a representation of geographic locations together with relationships between those locations. We define Directional Nearest-Neighbor graphs, prove a lemma asserting positive in-valence for all vertices and then prove strong connectivity for an important subclass of the graphs.

2. Definitions. Let $R_i(v)$ be the ith region of r equal divisions of the plane relative to the point v, as in Fig. 1. Let ball (v, w) be the interior region of the circle with center v and radius $d(v, w)$, where d is the standard Euclidean metric. For $v \neq w$, let $\text{dir}(v, w) = i$, where $w \in R_i(v)$. Define $\text{pie}(v, w)$ to be the intersection of $R_i(v)$ and ball (v, w), where $i = \text{dir}(v, w)$. Then a *Directional Nearest-Neighbor graph* with r divisions of the plane (DNN_r) is a directed graph for which the vertex set V is a finite subset of $\mathbb{R}^2$ and the edge set is $E = \{(v, w) \mid v, w \in V \text{ and } \exists\, u \in V \text{ s.t. } u \in \text{pie}(v, w)\}$.

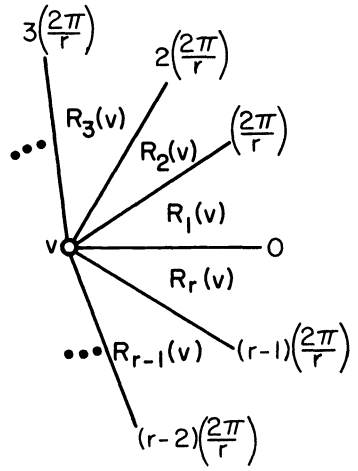


FIG. 1

The following results are for DNN_r graphs having $r = 4$. However, the properties also hold for DNN_{4k} graphs, $k > 0$, since every DNN_{4k} has a DNN_4 as a subgraph.

3. Strong connectivity of DNN_4 graphs. Let D_n denote a DNN_4 graph having n vertices and let $v^-(x)$ be the in-valence of a vertex x.

* Received by the editors March 25, 1981.

† Department of Computer and Information Science, Ohio State University, Columbus, Ohio 43210.
‡ Massachusetts Institute of Technology, Lincoln Laboratory, Lexington, Massachusetts 02173.

LEMMA. $v^-(x) > 0$ *for all* $x \in V(D_n)$, $n > 1$.

Proof. Let D_n be the smallest graph (with respect to n) containing a vertex p such that $v^-(p) = 0$. Then if any vertex is removed from D_n, the resulting D_{n-1} must have $v^-(x) > 0$ for all $x \in V(D_{n-1})$. Therefore, for every vertex $x_i \neq p$ in $V(D_n)$, there must be a vertex $x_c \neq p$ such that $(x_c, x_i) \in E(D_n)$ and $(x_c, p) \in E(D_n - \{x_i\})$. That is, there must be an edge (x_c, x_i) that is "changed" to (x_c, p) when x_i is removed. It will now be shown that the existence of such an x_c leads to a contradiction by choosing a particular x_i, namely x_m, defined as follows.

Let $M(x_i) = \max\{|x\text{-component of } x_i|, |y\text{-component of } x_i|\}$ and let x_m be such that $M(x_m)$ is a maximum over all $x_i \neq p$ in $V(D_n)$. Without loss of generality, assume that p has coordinates (0, 0), that $x_m = (-u, -v)$, $u, v > 0$, and that $v \geqq u$, as illustrated in Fig. 2. It will now be shown that x_c cannot lie in any of the indicated regions.

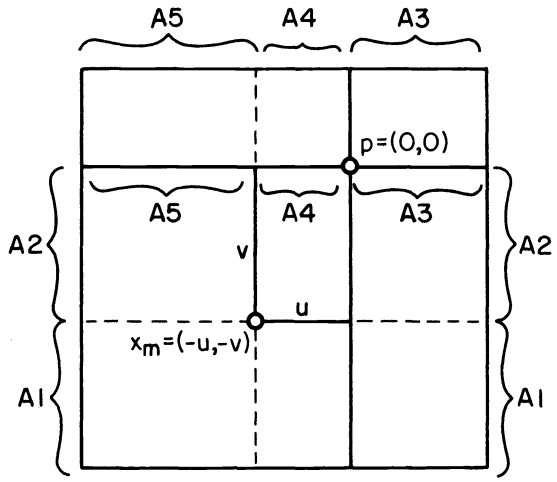


FIG. 2

Region A1. x_c cannot lie in this region, since x_m has the maximum component in any direction and, hence, in the y-direction.

Region A2. Any edge from an x_c in this region to x_m could not be changed to (x_c, p) when x_m is removed, since x_m and p are not in the same $R_i(x_c)$. Therefore, x_c cannot lie in this region.

Region A3. For any x_c in this region, pie(x_c, x_m) contains p. Hence edge (x_c, x_m) would not be in D_n and, therefore, x_c is not in Region A3.

Region A4. The argument for Region A2 also excludes x_c from Region A4.

Region A5. Suppose x_c is in this region. Then $x_c = (-s, t)$, $s, t > 0$, as illustrated in Fig. 3. Since $(x_c, x_m) \in E(D_n)$, $d(x_c, x_m) < d(x_c, p)$ must be the case. Consider the distance from x_c to x_m:

$$d(x_c, x_m) = ((-s+u)^2 + (t+v)^2)^{1/2}$$
$$= ((s^2+t^2) + (u^2 - 2su + v^2) + 2tv)^{1/2}.$$

Since v is the maximum component, $s \leqq v$, so that

$$u^2 - 2su + v^2 \geqq u^2 - 2vu + v^2 = (u-v)^2 \geqq 0$$

and t, $v > 0$ implies that $2tv > 0$. Therefore, $d(x_c, x_m) > (s^2+t^2)^{1/2} = d(x_c, p)$. But $d(x_c, x_m) < d(x_c, p)$. Therefore x_c cannot lie in Region A5. Thus x_c does not exist, contradicting the assumption $v^-(p) = 0$. Hence, the lemma. □

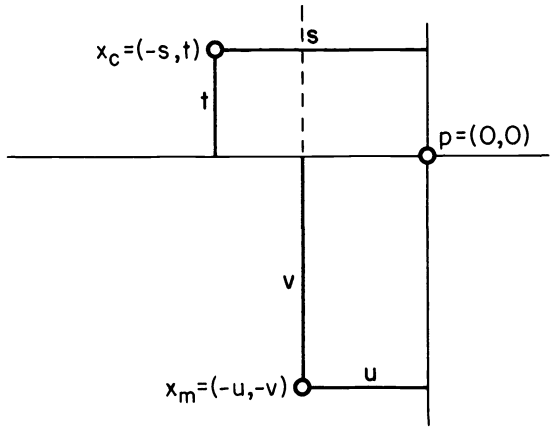


FIG. 3

THEOREM. *D_n is strongly connected.*

Proof. Let $v \sim w$ represent the existence of a directed path from v to w in D_n, and suppose D_n is the smallest graph (with respect to n) that is not strongly connected (i.e. $x \not\sim y$ for some x, $y \in V(D_n)$). Consider $S = \{s \in V(D_n) \mid x \sim s\}$ and $T = \{t \in V(D_n) \mid t \sim y\}$. It must be that $S \cap T = \varnothing$ or else S and T would have at least one vertex, u, in common such that $x \sim u$ and $u \sim y$ which would imply $x \sim y$—a contradiction. Thus no edges are directed from S to T.

Now consider the removal of $t \in T$, $t \neq y$, from D_n. That such a vertex must exist follows directly from the definition of T and the lemma. Since there are no edges from S to T, pie(s, t) contains at least one other vertex, where s is an arbitrary element of S. Let s' be closest to s in pie(s, t). Then, by definition of an edge, $(s, s') \in E(D_n)$. Furthermore, s' is in S (since $x \sim s$ and $s \sim s'$ imply $x \sim s'$). So t can be removed from D_n without changing any edges directed from elements of S. Thus S contains exactly the same vertices in $D_n - \{t\}$ as it did in D_n. Of course, since $y \notin S$, we have that $x \not\sim y$ in $D_n - \{t\}$, contradicting the assumption that D_n is the smallest graph that is not strongly connected. Hence, the theorem. □

SIAM J. ALG. DISC. METH.
Vol. 2, No. 4, December 1981

0196-5212/81/0204-0013 $01.00/0

THE RELIABILITY OF STANDBY SYSTEMS WITH A FAULTY SWITCH*

T. DOWNS† AND P. K. W. CHAN†

Abstract. In this paper explicit expressions are derived for the reliability of standby systems with a faulty switch. Three modes of switch malfunction are included, and the expressions apply to hot, warm and cold standby. The expressions are derived by finding the exponential of the state transition matrix using constituent matrices.

In order to increase the reliability of a unit within a system, standby redundancy is often used. One or more redundant units are attached to the "basic" unit and, when the basic unit fails, a redundant unit takes its place. Three types of standby redundancy are usually distinguished. In a system with *hot* standby, the redundant units are fully energized and are assumed to have the same failure distribution as the basic unit. *Cold* standby units are not supplied with power and it is assumed that they do not fail. *Warm* standby units are partially energized and, over any finite time interval, they have a smaller probability of failure than the basic unit.

The action of replacement of a failed basic unit by one of the redundant units is often carried out by an automatic switch. In practice, such switches are not perfectly reliable. The problem of switching unreliability has been considered by Gnedenko et al. [2]. They derived the Laplace transform for the reliability of a cold standby system whose switch was subject to three modes of malfunction. They did not proceed to invert the Laplace transform but pointed out that a "cumbersome sum of terms" would be obtained. It is the purpose of this paper to show that by making minor modifications to the switch model employed in [2] (these modifications being quite justifiable from a practical point of view) a relatively simple analytic form for the reliability function can be obtained, not only for the cold standby case but also for hot and warm standby.

The failure distributions of the basic and redundant units are assumed exponential with parameters α and μ, respectively. In the case of hot standby $\mu = \alpha$, for cold standby $\mu = 0$, for warm[1] standby $0 < \mu < \alpha$.

Three modes of switch malfunction are distinguished:

(i) *Ordinary failure.* When the switch fails ordinarily, it is unable to detect any faults in the basic unit and thus fails to switch to the next redundant unit when required. Such failures are assumed exponentially distributed with parameter γ.

(ii) *Clinging failure.* We assume that there is a probability $1-p$ that the switch can detect a failure in the basic unit but is unable to switch in the next redundant unit.

(iii) *False switching.* In this mode, the switch replaces the basic unit by one of the standbys when this is not required. We assume that the time between false switching is exponential with parameter ν.

The model employed in [2] allows clinging failure and false switching to take place simultaneously. From a practical viewpoint, this seems a most unlikely combination and is not allowed in our model. In addition, in [2] false switching was allowed to occur when no further standby units were available. This possibility is also excluded from our model. The failures of the units, as well as the various modes of switch malfunction, are

* Received by the editors December 6, 1978, and in final form March 24, 1980.

† Department of Electrical Engineering, University of Queensland, St. Lucia. Q. 4067, Australia.

[1] Note that we are considering a rather simple form of warm standby. More generally, the parameters of the standby units are not all the same.

all assumed independent. Finally, we assume that a standby unit which fails is immediately removed from the system.[2]

Let n be the total number of identical redundant units. Let (i, w) denote the state in which there are i "good" redundant units and the switch is in working order and let (i, f) denote the state in which there are i "good" redundant units and the switch has suffered an ordinary failure. The state transition diagram is given in Fig. 1. It is convenient to number the system states such that state $(n+1-k, w)$ is assigned the number k $(1 \leqq k \leqq n+1)$, state (i, f) is numbered $n+2$ and the failure state is numbered $n+3$.

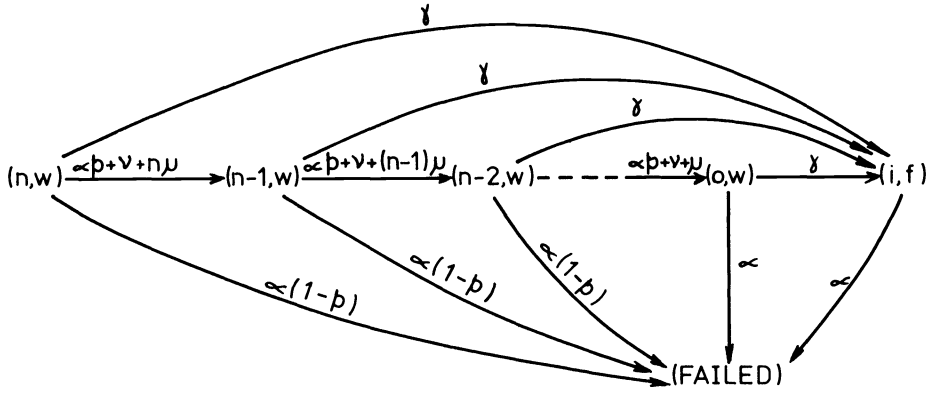


FIG. 1. *State transition diagram.*

Let $\underset{\sim}{P}(t)$ be an $(n+3)$-dimensional row vector whose ith entry, $p_i(t)$, is the probability that the system is in state i at time t. The state equation describing the process is

$$\frac{d}{dt}\underset{\sim}{P}(t) = \underset{\sim}{P}(t)\underset{\sim}{A}, \tag{1}$$

where $\underset{\sim}{A}$ is the $(n+3)\times(n+3)$ state transition matrix whose element $a_{ij}(i \neq j)$ is given by

$$a_{ij} = \lim_{\delta t \to 0}\left(\frac{\Pr[\text{system changes from state } i \text{ to } j \text{ in } (t, t+\delta t)]}{\delta t}\right) \tag{2}$$

and

$$a_{ii} = -\sum_{j \neq i} a_{ij}. \tag{3}$$

In our model, the transition matrix $\underset{\sim}{A}$ is given by

$$\underset{\sim}{A} = \tag{4}$$

$$\begin{bmatrix}
-(\alpha+\nu+n\mu+\gamma) & \alpha p+\nu+n\mu & 0 & \cdots & 0 & \gamma & \alpha(1-p) \\
0 & -[\alpha+\nu+(n-1)\mu+\gamma] & \alpha p+\nu+(n-1)\mu & \cdots & 0 & \gamma & \alpha(1-p) \\
0 & 0 & -[\alpha+\nu+(n-2)\mu+\gamma] & \cdots & 0 & \gamma & \alpha(1-p) \\
\cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdots & 0 & \gamma & \alpha(1-p) \\
\cdot & \cdot & \cdot & \cdots & \alpha p+\nu+\mu & \gamma & \alpha(1-p) \\
\cdot & \cdot & \cdot & \cdots & -(\alpha+\gamma) & \gamma & \alpha \\
0 & 0 & 0 & \cdots & 0 & -\alpha & \alpha \\
0 & 0 & 0 & \cdots & 0 & 0 & 0
\end{bmatrix}$$

[2] Note that this final assumption is not realistic in some practical situations.

If the initial state $\underset{\sim}{P}(0)$ is known, then $\underset{\sim}{P}(t)$ can be found from the state equation

$$\underset{\sim}{P}(t) = \underset{\sim}{P}(0) \exp(\underset{\sim}{A}t). \tag{5}$$

All units are assumed to be working at time $t = 0$, so that $\underset{\sim}{P}(0) = (1, 0, 0, \cdots, 0)$. Hence to find $\underset{\sim}{P}(t)$ it is only necessary to find the first row of the matrix $\exp(\underset{\sim}{A}t)$. In particular, $F(t)$ the cumulative failure distribution function is equal to $p_{n+3}(t)$, and is thus given by the $(1, n+3)$ element of $\exp(\underset{\sim}{A}t)$. Explicit expressions for $F(t)$ are given in the following theorems which are proved using a sequence of lemmas.

THEOREM 1. *For a system with n identical warm or hot standby units, the failure distribution $F(t)$ is given by*

$$F(t) = 1 + \sum_{i=1}^{n+2} z_{i0} \exp(\lambda_i t), \tag{6}$$

where λ_i is the ith *eigenvalue of $\underset{\sim}{A}$ (equal to the* ith *diagonal element of $\underset{\sim}{A}$), i.e.,*

$$\lambda_i = \begin{cases} -[(n+1-i)\mu + \alpha + \nu + \gamma] & \text{for } i \leqq n, \\ -(\alpha + \gamma) & \text{for } i = n+1, \\ -\alpha & \text{for } i = n+2, \end{cases} \tag{7}$$

and the coefficient z_{i0} is given by

$$z_{i0} = \frac{(-1)^{n-i}\alpha p}{(i-1)!(n-i)![(n-i+1)\mu + \gamma + \nu]} \prod_{\substack{j=1 \\ j \neq n-i+1}}^{n} \{j + (\alpha p + \nu)/\mu\} \quad \text{for } i \leqq n, \tag{8}$$

$$z_{n+1,0} = 0, \tag{9}$$

$$z_{n+2,0} = -\left[1 + \sum_{i=1}^{n} z_{i0}\right]. \tag{10}$$

THEOREM 2. *For a system with n cold standby units, the failure distribution $F(t)$ is given by*

$$F(t) = 1 + \sum_{i=0}^{n-1} z_{1i} t^i e^{-\lambda t} - (1 + z_{10}) e^{-\alpha t}, \tag{11}$$

where

$$\lambda = \alpha + \mu + \gamma \tag{12}$$

and

$$z_{1,n-k} = \frac{\alpha p (\alpha p + \nu)^{n-k}}{(n-k)!(\gamma + \nu)^k} \sum_{j=0}^{k-1} (\alpha p + \nu)^j (\gamma + \nu)^{k-j-1} \quad \text{for } 1 \leqq k \leqq n. \tag{13}$$

Our approach to the proof of the above theorems is based upon the following result, taken from Gantmacher [1].

For any function f which is analytic at the eigenvalues $\lambda_1, \lambda_2, \cdots \lambda_s$ of the matrix $\underset{\sim}{A}$, the fundamental formula for $f(\underset{\sim}{A})$ can be written as

$$f(\underset{\sim}{A}) = \sum_{k=1}^{s} \sum_{l=0}^{m_k - 1} f^{(l)}(\lambda_k) \underset{\sim}{Z}_{kl}, \tag{14}$$

where m_k is the multiplicity of the eigenvalue λ_k, $f^{(l)}$ is the lth derivative of f with respect to λ, and the matrices $\underset{\sim}{Z}_{kl}$, known as the constituent matrices or the components of the

matrix $\underset{\sim}{A}$, are independent of the function f and depend exclusively on $\underset{\sim}{A}$. In particular, for $f(\lambda) = \exp(\lambda t)$, $\exp(\underset{\sim}{A}t)$ is given by

$$\exp(\underset{\sim}{A}t) = \sum_{k=1}^{s} \sum_{l=0}^{m_k - 1} t^l \exp(\lambda_k t) \underset{\sim}{Z}_{kl}. \tag{15}$$

LEMMA 1. *For the matrix $\underset{\sim}{A}$ given in* (4), *let*

$$\underset{\sim}{M}_i = \underset{\sim}{A} + (\alpha + \nu + (n-i)\mu + \gamma)\underset{\sim}{I}. \tag{16}$$

Then the first row of the matrix

$$\underset{\sim}{P} = \underset{\sim}{M}_0 \underset{\sim}{M}_1 \underset{\sim}{M}_2 \cdots \underset{\sim}{M}_{i-2} \tag{17}$$

for $2 \leqq i \leqq n-1$ has only three nonzero entries, these being entries $(1, i)$, $(1, n+2)$ and $(1, n+3)$.

Furthermore, the $(1, i)$ entry is given by

$$p_{1i} = \prod_{j=0}^{i-2} [\alpha p + \nu + (n-j)\mu]. \tag{18}$$

Proof. By multiplication, this is true for $i = 3$. We will now assume it to be true for some other i in the range $(4 \leqq i \leqq n-2)$ and show that such an assumption implies that it is true for $i+1$.

Consider the postmultiplication of $\underset{\sim}{P}$ by $\underset{\sim}{M}_{i-1}$. Note that $\underset{\sim}{M}_{i-1}$ is upper triangular and that in its ith row there are only three nonzero entries, in positions $(i, i+1)$, $(i, n+2)$ and $(i, n+3)$. This immediately implies that the first row of the matrix $\underset{\sim}{P}\underset{\sim}{M}_{i-1}$ will have only three nonzero entries, in positions $(1, i+1)$, $(1, n+2)$ and $(1, n+3)$, as required. Furthermore, since entry $(i, i+1)$ of $\underset{\sim}{M}_{i-1}$ is equal to $(\alpha p + \nu + (n+1-i)\mu)$, the lemma follows. □

LEMMA 2. *Column $n+3$ of the matrix*

$$\underset{\sim}{Q} = \underset{\sim}{M}_i \underset{\sim}{M}_{i+1} \underset{\sim}{M}_{i+2} \cdots \underset{\sim}{M}_{n-1} [\underset{\sim}{A} + \alpha \underset{\sim}{I}] \underset{\sim}{A} \tag{19}$$

for $1 \leqq i \leqq n-1$ has entries given by

$$q_{k,n+3} = \begin{cases} \alpha p[\gamma + \alpha(1-p)] \prod_{r=1}^{n-i} (\alpha p + \nu + r\mu), & 1 \leqq k < i, \\ \alpha p[\alpha + (n-i+1)\mu + \gamma + \nu] \prod_{r=1}^{n-i} (\alpha p + \nu + r\mu), & k = i, \\ 0, & i < k \leqq n+3. \end{cases} \tag{20}$$

Proof. By direct multiplication it can be shown that the lemma is true for $i = n-1$. We will now assume it to be true for some other i in the range $2 \leqq i \leqq n-2$ and show that such an assumption implies that it is true for $i-1$.

Consider the premultiplication of $\underset{\sim}{Q}$ by $\underset{\sim}{M}_{i-1}$. Denote by $\underline{q}^j$ the jth column of $\underset{\sim}{Q}$ and by $\underline{m}^k$ the kth row of $\underset{\sim}{M}_{i-1}$. In addition, denote the ith entry of $\underline{m}^k$ by m_{ki}. Now note that for $1 \leqq k \leqq i-1$, $\underline{m}^k$ has only four nonzero entries, viz. $m_{k,k} = (k-i)\mu$, $m_{k,k+1} = \alpha p + \nu + (n-k+1)\mu$, $m_{k,n+2} = \gamma$ and $m_{k,n+3} = \alpha(1-p)$. Thus postmultiplication of this row by $\underline{q}^{n+3}$ gives, for $k \leqq i-2$,

$$\underline{m}^k \underline{q}^{n+3} = \alpha p[\gamma + \alpha(1-p)] \prod_{r=1}^{n-i+1} (\alpha p + \nu + r\mu)$$

as required.

Furthermore, for $k = i-1$

$$\underline{m}^{i-1}\underline{q}^{n+3} = \alpha p\{[\gamma + \alpha(1-p)](-\mu) + [\alpha + (n-i+1)\mu + \gamma + \nu][\alpha p + \nu + (n-i)\mu]\} \prod_{r=1}^{n-i} (\alpha p + \nu + r\mu)$$

$$= \alpha p[\alpha + (n-i+2)\mu + \gamma + \nu] \prod_{r=1}^{n-i+1} (\alpha p + \nu + r\mu),$$

as required. Now $\underline{m}^i$ has only three nonzero entries, viz. $m_{i,i+1}$, $m_{i,n+2}$ and $m_{i,n+3}$ and hence $\underline{m}^i\underline{q}^{n+3} = 0$. And since the entries of $\underline{m}^k\underline{q}^{n+3}$ are zero for $k > i$ by assumption, the lemma is proved. □

We are now in a position to prove Theorem 1. For hot or warm standby systems, the eigenvalues of the matrix $\underset{\sim}{A}$ are all distinct and hence (16) reduces to

$$\exp(\underset{\sim}{A}t) = \sum_{i=1}^{n+3} \exp(\lambda_i t)\underset{\sim}{Z}_{i0}. \tag{21}$$

The constituent matrices $\underset{\sim}{Z}_{i0}$ may be computed by inserting in the fundamental formula (15) appropriate trial functions $f(\lambda)$. Since only the $(1, n+3)$ element of $\exp(\underset{\sim}{A}t)$ is required, it is only necessary to calculate the $(1, n+3)$ elements of the matrices $\underset{\sim}{Z}_{i0}$. In order to keep the notation simple, let z_{i0} be the $(1, n+3)$ element of the matrix $\underset{\sim}{Z}_{i0}$. Then we have

$$F(t) = \sum_{i=1}^{n+3} \exp(\lambda_i t) z_{i0}. \tag{22}$$

For $z_{n+1,0}$, let the trial function $f(\lambda)$ be

$$f(\lambda) = \prod_{\substack{i=1 \\ i \neq n+1}}^{n+3} (\lambda + \lambda_i).$$

From the fundamental formula (15)

$$\underset{\sim}{Z}_{n+1,0} \prod_{\substack{i=1 \\ i \neq n+1}}^{n+3} (\lambda_i - \alpha - \gamma) = \underset{\sim}{M}_0\underset{\sim}{M}_1 \cdots \underset{\sim}{M}_{n-1}[\underset{\sim}{A} + \alpha\underset{\sim}{I}]\underset{\sim}{A} \tag{23}$$

$$= \underset{\sim}{P}\underset{\sim}{M}_{i-1}\underset{\sim}{Q}.$$

From Lemmas 1 and 2, the $(1, n+3)$ element of the matrix $\underset{\sim}{P}\underset{\sim}{M}_{i-1}\underset{\sim}{Q}$ is zero, and since the product on the left side of (23) is nonzero, $z_{n+1,0}$ must be equal to zero.

To find z_{k0} for $k \leqq n$, let the trial function $f(\lambda)$ be

$$f(\lambda) = \prod_{\substack{i=1 \\ i \neq k, n+1}}^{n+3} (\lambda + \lambda_i). \tag{24}$$

Substitution into (15) gives

$$\underset{\sim}{Z}_{k0} = \underset{\sim}{P}\underset{\sim}{Q}/\{(i-1)!\,(n-i)!\,\mu^{n-1}[(n+1-i)\mu + \nu + \gamma][\alpha + \nu + (n+1-i)\mu + \gamma]\}.$$

Using Lemmas (1) and (2), this yields (8).

The remaining z_{i0} (viz. $z_{n+2,0}$ and $z_{n+3,0}$) could be found by use of trial functions in a similar fashion, but this is not necessary since consideration of (22) as $t \to \infty$ shows that

$z_{n+3,0}=1$ (since $\lambda_{n+3}=0$) and letting $t\to 0$, (22) implies

$$z_{n+2,0}=-\sum_{\substack{i=1\\ i\neq n+2}}^{n+3} z_{i0},$$

which gives (10). □

Theorem 2 applies to a limiting case (as $\mu\to 0$) of the situation covered by Theorem 1 but cannot be easily obtained (by substituting $\mu=0$) from Theorem 1. For this reason we derive Theorem 2 separately. For a system with n cold standby units, the eigenvalue $\lambda_1=-(\alpha+\nu+\gamma)$ is of multiplicity n and the other eigenvalues are distinct. Hence the fundamental formula reduces to

$$\begin{aligned}f(\underset{\sim}{A})=f(\lambda_1)\underset{\sim}{Z}_{10}+f^{(1)}(\lambda_1)\underset{\sim}{Z}_{11}+\cdots+f^{(n-1)}(\lambda_1)\underset{\sim}{Z}_{1,n-1}\\ +f(\lambda_2)\underset{\sim}{Z}_{20}+f(\lambda_3)\underset{\sim}{Z}_{30}+f(\lambda_4)\underset{\sim}{Z}_{40}.\end{aligned}\tag{25}$$

In particular, $\exp(\underset{\sim}{A}t)$ is given by

$$\begin{aligned}\exp(\underset{\sim}{A}t)=e^{\lambda_1 t}\underset{\sim}{Z}_{10}+te^{\lambda_1 t}\underset{\sim}{Z}_{11}+\cdots+t^{n-1}e^{\lambda_1 t}\underset{\sim}{Z}_{1,n-1}\\ +e^{\lambda_2 t}\cdot\underset{\sim}{Z}_{20}+e^{\lambda_3 t}\underset{\sim}{Z}_{30}+e^{\lambda_4 t}\underset{\sim}{Z}_{40}\end{aligned}\tag{26}$$

As before, to find $F(t)$ we only require the $(1,n+3)$ entry z_{ij}, of each of the matrices $\underset{\sim}{Z}_{ij}$.

To find z_{1j} we employ the trial function

$$f_j(\lambda)=(\lambda+\alpha+\nu+\gamma)^j(\lambda+\alpha)\lambda,\tag{27}$$

which requires us to compute the $(1,n+3)$ entry of the matrix $\underset{\sim}{R}^{(j)}$, given by

$$\underset{\sim}{R}^{(j)}=[\underset{\sim}{A}+(\alpha+\nu+\gamma)\underset{\sim}{I}]^j[\underset{\sim}{A}+\alpha\underset{\sim}{I}]\underset{\sim}{A}.\tag{28}$$

Denote by $r_{lm}^{(j)}$ element (l,m) of the matrix $\underset{\sim}{R}^{(j)}$.

LEMMA 3. *The* $(1,n+3)$ *element of* $\underset{\sim}{R}^{(j)}$ *is given by*

$$r_{1,n+3}^{(j)}=\begin{cases}\alpha p[\gamma+\alpha(1-p)](\alpha p+\nu)^j, & 0\leqq j<n-1,\\ \alpha p(\gamma+\alpha+\nu)(\alpha p+\nu)^j, & j=n-1,\\ 0, & j=n.\end{cases}\tag{29}$$

Proof. For $j\leqq n$, the jth row of $\underset{\sim}{A}+\alpha\underset{\sim}{I}$ has four nonzero entries, viz. (j,j), $(j,j+1)$, $(j,n+2)$ and $(j,n+3)$ and in particular, element $(j,j+1)$ is equal to $\alpha p+\nu$, element $(j,n+2)$ is equal to γ and element $(j,n+3)$ is equal to $\alpha(1-p)$. The first n entries in column $n+3$ of $\underset{\sim}{A}$ are equal to $\alpha(1-p)$, elements $(n+1,n+3)$ and $(n+2,n+3)$ are equal to α and element $(n+3,n+3)$ is equal to zero. Thus $r_{i,n+3}^{(0)}=\alpha p[\gamma+\alpha(1-p)]$ for $i<n$ and is equal to $\alpha p(\alpha+\mu+\nu+\gamma)$ for $i=n$. By similar reasoning, it is easy to see that $r_{i,n+3}^{(0)}=0$ for $i>n$.

The first n columns of $\underset{\sim}{A}+(\alpha+\nu+\gamma)\underset{\sim}{I}$ are null except for entries $(i,i+1)$ which are equal to $\alpha p+\nu$ (for $1\leqq i\leqq n-1$). Thus

$$r_{i,n+3}^{(1)}=\begin{cases}\alpha p[\gamma+\alpha(1-p)](\alpha p+\nu), & i<n-1,\\ \alpha p[\gamma+\alpha+\nu](\alpha p+\nu), & i=n-1,\\ 0, & i\geqq n,\end{cases}\tag{30}$$

and the lemma follows by repeated multiplication. □

In order to proceed we also require certain derivatives of the trial function (27). Our needs are satisfied by the following lemma, which can be proved by a simple inductive argument.

LEMMA 4. *For $l \leqq i$, the lth derivative of the testing function $f_j(\lambda)$ in* (27) *is given by*

$$f_j^{(l)}(\lambda) = j_{(l)}\lambda(\lambda+\alpha)(\lambda+\alpha+\nu+\gamma)^{j-l} + lj_{(l-1)}(2\lambda+\alpha)(\lambda+\alpha+\nu+\gamma)^{j-l+1} + l(l-1)j_{(l-2)}(\lambda+\alpha+\nu+\gamma)^{j-l+2} \tag{31}$$

where $j_{(l)}$ represents the falling factorial $j(j-1)\cdots(j-l+1)$.

In order to prove Theorem 2, we introduce one further lemma. From (26), for a system with n cold standby units, the failure distribution $F(t)$ is given by

$$F(t) = e^{\lambda_1 t}z_{10} + te^{\lambda_1 t}z_{11} + \cdots + t^{n-1}e^{\lambda_1 t}z_{1,n-1} + e^{\lambda_2 t}z_{20} + e^{\lambda_3 t}z_{30} + e^{\lambda_4 t}z_{40}. \tag{32}$$

LEMMA 5. *The coefficient z_{20} is equal to zero and*

$$z_{1,n-k} = \frac{\alpha p(\alpha p+\nu)^{n-k}}{(n-k)!(\gamma+\nu)^k}\sum_{i=0}^{k-1}(\alpha p+\nu)^i(\gamma+\nu)^{k-i-1} \quad \text{for } 1 \leqq k \leqq n. \tag{33}$$

Proof. Application of the testing function $f_n(\lambda) = (\lambda+\alpha+\nu+\gamma)^n(\lambda+\alpha)\lambda$, in conjunction with Lemmas 4 and 5, gives $z_{20} = 0$.

To find $z_{1,n-1}$, use the testing function $f_{n-1}(\lambda)$ given in (27). From Lemmas 4 and 5, we have

$$z_{1,n-1} = \frac{\alpha p(\alpha p+\nu)^{n-1}}{(n-1)!(\gamma+\nu)}$$

showing that the lemma is true for $k = 1$. Similarly, using testing function $f_{n-2}(\lambda)$ it can be shown that the lemma is true for $k = 2$. Now assume that it is true for $z_{1,n-k}(3 \leqq k \leqq n-1)$ and show that it is true for $z_{1,n-k-1}$. To find $z_{1,n-k-1}$, use testing function $f_{n-k-1}(\lambda)$ from (27), and using lemmas 3 and 4, we obtain

$$(n-k-1)!(\alpha+\gamma+\nu)(\gamma+\nu)z_{1,n-k-1} + (n-k)!(\alpha+2\gamma+2\nu)z_{1,n-k} + (n-k+1)!z_{1,n-k+1} = \alpha p[\gamma+\alpha(1-p)](\alpha p+\nu)^{n-k-1}. \tag{34}$$

On substituting $z_{1,n-k}$ and $z_{1,n-k+1}$ into (34) and collecting terms, we have

$$z_{1,n-k-1} = \frac{\alpha p(\alpha p+\nu)^{n-k-1}}{(n-k-1)!(\alpha+\gamma+\nu)(\gamma+\nu)^{k+1}}\{T_1+T_2+T_3\}, \tag{35}$$

where

$$T_1 = [\gamma+\alpha(1-p)](\gamma+\nu)^k, \tag{36}$$

$$T_2 = (\alpha p+\nu)(\alpha+2\gamma+2\nu)\sum_{i=0}^{k-1}(\alpha p+\nu)^i(\gamma+\nu)^{k-i-1}, \tag{37}$$

$$T_3 = -(\alpha p+\nu)^2\sum_{i=0}^{k-2}(\alpha p+\nu)^i(\gamma+\nu)^{k-i-1}. \tag{38}$$

After some algebraic manipulation, this reduces to

$$z_{1,n-k-1} = \frac{\alpha p(\alpha p+\nu)^{n-k-1}}{(n-k-1)!(\gamma+\nu)^{k+1}}\sum_{i=0}^{k}(\alpha p+\nu)^i(\alpha+\nu)^{k-i},$$

and the lemma follows. □

By considering (32) as $t \to \infty$ and $t \to 0$, we find $z_{40} = 1$ and $z_{30} = -(1 + z_{10})$. These results, together with Lemma 5, prove Theorem 2. □

REFERENCES

[1] F. R. GANTMACHER, *The Theory of Matrices*, Vol. I, Chelsea, New York, 1959.

[2] B. V. GNEDENKO, YU. K. BELYAYEV AND A. D. SOLOVYEV, *Mathematical Methods of Reliability Theory*, Academic Press, New York, 1969.